Observation Oriented Modeling

Analysis of Cause in the Behavioral Sciences

James W. Grice

OBSERVATION ORIENTED MODELING

ELSEVIER science & technology books

Companion Web Site:
http://booksite.academicpress.com/grice/oom/

Observation Oriented Modeling by James W. Grice

Resources:

- Foreword by Paul Barrett
- Book - Table of Contents
- How to Use OOM Software (Technical Manual)
- OOM Example Data Files
- OOM Software Program

ELSEVIER

ACADEMIC
PRESS

To adopt this book for course use, visit http://textbooks.elsevier.com

OBSERVATION ORIENTED MODELING

Analysis of Cause in the Behavioral Sciences

JAMES W. GRICE, Ph.D.
Oklahoma State University, Stillwater, OK, USA

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
32 Jamestown Road, London NW1 7BY, UK
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA

First edition 2011

Copyright © 2011 Elsevier Inc. All rights reserved

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+ 44) (0) 1865 843830; fax (+44) (0) 1865 853333; email: permissions@elsevier.com. Alternatively, visit the Science and Technology Books website at www.elsevierdirect.com/rights for further information

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978-0-12-385194-9

For information on all Academic Press publications visit our website at www.elsevierdirect.com

Typeset by TNQ Books and Journals

Printed and bound in United States of America

11 12 13 14 10 9 8 7 6 5 4 3 2 1

Working together to grow libraries in developing countries
www.elsevier.com | www.bookaid.org | www.sabre.org
ELSEVIER BOOK AID International Sabre Foundation

CONTENTS

Foreword by Paul Barrett *ix*
Acknowledgments *xiii*

1. Introduction **1**

Metaphysical and Methodological Errors 1
Observation Oriented Modeling and Philosophical Realism 3

2. Data at Its Core **9**

Introduction 9
Deep Structure 10
Multiple Observations 20
Deep Structure Addition and Subtraction 22
Deep Structure Multiplication and Division 24
Logical Operations and Deep Structure 25

3. Rotating Deep Structures **29**

Introduction 29
Procrustes Rotation 29
Binary Procrustes Rotation 36

4. Modeling with Deep Structures **45**

Introduction 45
A Simple Experiment 46
Modeling the Experiment 47
Graphs Both Old and New 53
A Probability Statistic 57
Concluding Models and Summary 59

5. Statistics and Null Hypothesis Significance Testing **65**

Introduction 65
The Force of History 66
Null Hypothesis Significance Testing 74

Turning the Tide with Observation Oriented Modeling 83
Hybridized NHST 84

6. Modeling and Inferential Statistics 87

Introduction 87
Modeling in Science 88
The Bystander Effect Model 94
Estimating Population Parameters in NHST 99
Modeling Persons, Testing Aggregates 104
Observation Oriented Modeling 108

7. Modeling and Effect Sizes 121

Introduction 121
Aggregate Effects 122
Differences Between Means 123
Correlation and Shared Variance 136

8. Measurement and Additive Structures 143

Introduction 143
Continuous Quantitative Measurement 145
Measurement as Knowing Quantity 151
Observation Oriented Modeling 158
Additive Models 166
Measurement Error 169
Latent Variables 175

9. Cause and Effect 183

Introduction 183
Causality and Causation 184
The Randomized Controlled Experiment 187
Mediation Models 190
Person-Centered Mediation Analysis 195
Logical Hypothesis Testing 201
Final, Formal, and Efficient Causes 206
Transition to the Four Causes 214

10. Coda **219**

Introduction 219
Seven Principles of Observation Oriented Modeling 221
Summary 230

References *231*
Index *239*

A companion website to support this book can be found at:
http://booksite.academicpress.com/grice/oom/

This page intentionally left blank

FOREWORD

This is a book that you will find difficult to stop reading, and it is a book that you will not easily forget. There are facts in this book about causal analysis, measurement, and statistical methodology that will cause you to rethink much of what you assumed or were told was "best practice." It is also an exciting book because of what it promises the social scientist—a pioneering venture into the undiscovered country of a 21st-century psychological science.

It is a remarkable book—not because of the scholarship, philosophical discourse, logic, and methodological expertise clearly on display but, rather, because it signals a return to the fundamentals of a science of psychology. That science has been distorted over many decades by those who think that mostly inaccurate statistical generalizations that rely on a host of dubious and untested assumptions can replace scientific knowledge acquired by investigators who espouse explanatory accuracy over statistical model fit. James Grice has exposed this fallacy with careful and comprehensive philosophical and logical analysis, explaining why the current obsession with statistical modeling has led to almost complete ideational and scientific stagnation within the area of "quantitative" psychology.

However, where many might have been content with the argument and discourse alone, James has taken the next step. And this is some step. Within this book and accompanying software is what I truly believe to be the next-generation analysis methodology for psychological science.

Observation oriented modeling is not just another statistical or data analysis methodology like structural equation modeling or analysis of variance. Rather, it is a way of testing theories, ideas, and hypotheses about data that delivers usable and readily understandable facts about the observational data themselves, not idealized or sanitized versions of them. That data may be qualitative or quantitative, where relations between data structures may be quantities, orders, or sets of qualitative rule relations. Observation oriented modeling also delivers direct and rather obvious estimates of accuracy—not effect sizes, *p* levels, or *r*-squares, but much more straightforward evaluation of the precise consequences specified by a theory or hypothesis. The software provided for this purpose is unique, simple to use, and effective.

When you reach the end of this book, you will have begun your first faltering steps into a new kind of science—one that honors the four causes of Aristotle, the integrative science of Thomas Aquinas, and the unifying philosophy of William Wallace. You will also understand why current practice must change, how it must change, and the new methodology for investigative causal analysis by which change can be made a reality. It is not often that such a unique, field-changing opportunity is presented to a research community. I urge you to grasp this one with both hands, digest it fully, and then forge your own path into that undiscovered country.

—Paul Barrett
Auckland, New Zealand

For my parents and for my teachers—with special gratitude to Ellen Gadd, Vernon Polly, William Chambers, Richard Harris, and Harold Delaney

This page intentionally left blank

ACKNOWLEDGMENTS

I thank my beautiful wife, Melissa, and my three wonderful children—Jacob, Elizabeth, and Samuel—for their loving support and encouragement while I was completing this book and the Observation Oriented Modeling software. I also thank my current graduate students—Stefanie Badzinski, Jim Anderson, and Erika Brown—for proofreading and discussing this book as it was being written. They are, as William Chambers was fond of saying, "a pack of intellectuals" of whom I am most proud. Jude Dougherty and Peter Redpath read particular chapters and sections with a philosophical emphasis. I greatly appreciate the time they devoted to this work, and I of course accept full responsibility for any philosophical or historical errors that may remain. I also thank Nikki Levy and Elsevier for their support of this book and the accompanying materials. It is an honor to have the aid of such an experienced editor and publishing company. The faculty and administration at Oklahoma State University deserve a special thanks as well for their generosity in offering a sabbatical leave for the fall of 2009 during which time most of this book was written. Finally, I cannot write enough to express my gratitude to Paul Barrett, a true scholar with the requisite critical mind. His understanding of the psychometric issues covered in this book is deeper than my own, and his knowledge of the modern literature in statistics, testing, and assessment is truly remarkable. He brought to my attention, during the past 10 years, many of the scholarly articles cited in this book.

This page intentionally left blank

CHAPTER 1

Introduction

Contents

Metaphysical and Methodological Errors 1
Observation Oriented Modeling and Philosophical Realism 3

METAPHYSICAL AND METHODOLOGICAL ERRORS

Mortimer wrote candidly of his experiences as a young psychologist on his way to earning a doctoral degree from Columbia University in the 1920s. He described how psychology had been divided into various "schools of thought" and how he learned to recite the dogmas of these diverse schools. Through careful study, he learned to tell the difference between behaviorists, structuralists, Freudians, Gestalters, and Jungians—not because they had made significantly different discoveries regarding the human psyche but, rather, because they had developed their own "point of view." He also studied human physiology, endocrinology, and neuroanatomy, and he was schooled in the various techniques of psychological experimentation. In one study pertaining to emotions, he would carefully measure pupillary responses as his colleague lowered a boa constrictor onto the head of a wary participant. Yet, upon completing his dissertation, Mortimer was confronted with a frightening reality:

> *I could not tell my students, my colleagues, or myself, what psychology was about, what its fundamental principles were, or what was the theoretical significance of all the data and findings that thousands of young men like myself had been collecting and assorting ever since the Ph.D. industry and the research foundations had encouraged such labors.*

The astute reader will recognize this student of psychology as Mortimer J. Adler, one of the most visible and controversial philosophers of the 20th century (Adler, 1941, p. ix). He wrote of his experience with psychology in 1941, and nearly 70 years later his words accurately describe an academic discipline adrift in diversity with no apparent hope for real unity. From Adler's perspective, psychology lost its way through a series of metaphysical errors made early in its history. Other distinguished scholars, such as Rom Harré, Daniel Robinson, and Joseph Rychlak, have made similar arguments, noting that in the rush to divorce itself from philosophy,

Observation Oriented Modeling
ISBN 978-0-12-385194-9, Doi:10.1016/B978-0-12-385194-9.10001-5
© 2011 Elsevier Inc.
All rights reserved.

psychology fell into a number of metaphysical traps from which it has yet to fully recover (Harré & Secord, 1973; Robinson, 1986; Rychlak, 1988).

Many of the metaphysical difficulties faced by psychology pertain specifically to its prevailing research tradition and whether or not the premises underlying this tradition have been examined sufficiently. Psychology is in a unique position among the sciences because the object of study (a human person) is also the subject of study. A psychologist, for instance, who concludes from her research that individuals cannot view factual evidence in an unbiased manner must wonder if this conclusion is itself the result of her own biases. Such quandaries are not new to psychology, yet its prevailing methodology has been derived almost exclusively from a philosophical position (i.e., positivism) that denies the significance of the subject–object dialectic. This fact raises the specter of an important question: Are the prevailing research methods capable of yielding an authentic science of psychology? Asked in more specific forms, is significance testing the appropriate tool for evaluating data? Is the randomized controlled experiment sufficient for determining causality? Are parametric statistics appropriate for the attributes studied by psychologists? Such questions have been a nagging thorn in the side of psychology because they are routinely answered in the negative. In fact, more than a few prominent researchers in the field have publicly denounced the prevailing research methodology. For instance, in a book chapter titled "What's Wrong with Psychology, Anyway?" David Lykken (1991) stated plainly yet forcefully that psychology's research tradition is fundamentally flawed, describing most grant applications, submitted manuscripts, and published research as simply "bad." He also indicted the majority of published psychological studies as nonreplicable and as yielding very little in the way of cumulative knowledge. Interestingly, Lykken was simply sharpening the same criticisms made by Paul Meehl nearly 15 years earlier (Meehl, 1978). Jacob Cohen offered a more precise critique of psychology's prevailing research tradition, arguing that the primary tool for judging the outcome of any given study (i.e., the significance test) actually impedes the advance of psychology as a science. He stated, "I argue herein that NHST [null hypothesis significance testing] has not only failed to support the advance of psychology as a science but also has seriously impeded it" (Cohen, 1994, p. 997). Like Lykken, Cohen was simply restating a valid criticism made nearly 30 years earlier, this time by David Bakan (1967):

> *I will attempt to show that the test of significance does not provide the information concerning psychological phenomena characteristically attributed to it;*

> *and that, furthermore, a great deal of mischief has been associated with its use. If the test of significance does not yield the expected information concerning the psychological phenomena under investigation, we may well speak of a crisis; for then a good deal of the research of the last several decades must be questioned.*
>
> ***(p. 2)***

In a similar vein, Benjamin Wright lamented the failure of social scientists to heed previous arguments regarding the importance of establishing the quantitative structure of measured attributes. He concluded, "That is why so much social science has turned out to be no more than transient descriptions of never-to-be reencountered situations, easy to contradict with almost any replication" (Wright, 1997, p. 35). Finally, in two organizational efforts, the American Psychological Association (APA) commissioned a task force to offer recommendations for reforming the statistical reporting practices of psychologists, and the National Institutes of Mental Health (NIMH) gathered the editors of 24 scientific journals to endorse a statement calling for radical change in the prevailing research tradition of psychology (NIMH, 2000):

> *We believe that traditional, variable-oriented, sample-based research strategies and data analytic techniques alone cannot reveal the complex causal processes that likely give rise to normal and abnormal behavior among different children and adolescents. To a large extent, the predominant methods of our social and psychological sciences have valued quantitative approaches over all others, to the exclusion of methods which might clarify the ecological context of behavioral and social phenomena.*
>
> ***(p. 66)***[1]

Despite the efforts of these two organizations and those of the prominent and respected academics mentioned previously, no substantial, lasting change in psychology's research methodology has occurred.

OBSERVATION ORIENTED MODELING AND PHILOSOPHICAL REALISM

Given that the issues plaguing psychology's research tradition are at their root metaphysical, any valid way forward must be accompanied by a fundamental shift in philosophy; and it is on the basis of this premise that observation oriented modeling is herein presented as an entirely new way of conceptualizing and analyzing psychological data. As a novel set of methods, observation oriented modeling seeks to explain patterns of observations in terms of their causal structure. This approach stands in stark contrast to the

[1] The conclusions from the APA's task force were published in Wilkinson et al. (1999).

prevailing research strategy that is centered around variable modeling and the estimation of parameters (e.g., means, variances, and proportions) for populations of events that often exist only in theory. By focusing on causal structure, observation oriented modeling aligns itself with philosophical realism and stands in opposition to the positivism that produced psychology's current research tradition.

Philosophical realism has taken on a variety of forms over the course of history, but it nonetheless represents a continuous line of thought that can be traced all the way back to Plato and Aristotle. It has recently gained a strong foothold in the relatively young philosophy of science discipline and has found a number of prominent voices among physical and social scientists (e.g., see Bhaskar, 1975; Harré, 1987; Manicas, 2006). In general terms, philosophical realism holds (1) there is a world of real existence that is not made or constructed by humanity, (2) a person can know this existing reality in which he or she is an active participant, and (3) such knowledge is the most reliable guide to individual and social conduct (Wild, 1948).[2] These tenets may seem to represent common sense, and indeed realist philosophy is sometimes described as "reasoned common sense." It is also sometimes said that all scientists are realists in the laboratory, but most adopt idealism when participating in armchair philosophy and subjectivism when confronting questions of morality. Regardless, philosophical realism provides a number of distinct advantages over the positivism that has given rise to the prevailing research tradition in psychology. These advantages will be made manifest in subsequent chapters as observation oriented modeling is introduced, but a number of prefatory thoughts will greatly aid its introduction.

For the realist, the subject–object dialectic is valid and has merit because it means that a person can come to know the things of nature with increasing clarity and depth. For the moderate realist, following Aristotle and Thomas Aquinas, an act of knowing occurs when a person (subject) intentionally possesses the immaterial form of a thing (the object), and this possession involves both sensation and intellection.[3] When presented with a red book, for instance, a person is, through his senses, immediately aware

[2] Daniel Sullivan also provides a brief and perhaps more enjoyable introduction to philosophical realism in *An Introduction to Philosophy: Perennial Principles of the Classical Realist Tradition* (1957/2009).

[3] The term "moderate realism" is taken from Wallace (1983). It is meant to convey realism in the tradition of Aristotle and Thomas Aquinas. Wallace exemplifies this viewpoint in his book. See also Wallace (1996).

of the book as an individual existing thing; and at the same time his intellect abstracts from the red book the universal concepts "red" and "book." These concepts exist within the mind but also in some way within the book as well. Such real concepts are therefore not independent inventions of the mind, although they can be combined, separated, and manipulated in various ways within the mind; and they can be evaluated with other abstract concepts that exist solely within the mind, such as logical concepts and certain mathematical concepts (e.g., zero and null set) (Wallace, 1996, p. 255). New formulations of concepts can moreover be judged for their conformity with the things of nature, such as when a psychologist posits memory as separable into long-term and short-term components and then devises experiments to test these conceptualizations.

A consequence of this position is that a scientist can know something of the "things themselves," and it is perfectly legitimate to speak of the powers and properties of the things of nature. Although modern academics find it fashionable to quote Molière's clever criticism of Aristotle's notions of powers and properties, they run the risk of undercutting the possibility of scientific knowledge (an example can be found in Gigerenzer, 2009). Truly, it is circular to argue that opium causes drowsiness because of its dormative properties, but it is a graver mistake still to assume that caffeine or some other compound can all equally be used to explain drowsiness. This is simply not the case because there is something in the nature of opium that plays a causal role in drowsiness and that makes it different from caffeine. Scientists indeed depend on these different natures, at least implicitly, when making general and predictive statements about these very compounds. A realist will furthermore hold that powers and properties can be discovered and made known through the discernment of patterns in normal experience or in the special experience that constitutes a scientific experiment:

> *What are best termed the "abstract sciences" aim at an understanding of the fundamental processes of nature. Such inquiry may be motivated by discerning a pattern, but not all patterns will be of concern. Indeed, patterns which emerge from experimentally generated data, e.g., the results of Lavoisier's painstaking use of the chemical balance, are of high importance.*
>
> ***(Manicas, 2006, p. 25)***[4]

[4] Rom Harré (1986, p. 35) also states, "Theories are seen as solutions to a peculiar style of problem: namely, 'Why is it that the patterns of phenomena are the way they are?' A theory answers this question by supplying an account of the constitution and behavior of those things whose interactions with each other are responsible for the manifested patterns of behavior."

The concept of pattern also brings to the forefront other important concepts, such as whole–part relations, unity, and integration, all of which figure prominently in discussions of observation oriented modeling in subsequent chapters.

Presuming things to have particular powers because of their natures brings causality to the center of scientific reasoning as well. Consistent with the notion of pattern described previously, particular causes normally entail particular effects, and it is the primary goal of the scientist to create a model that explains these ordered (i.e., patterned) structures and processes. The models of the atom or human cell are excellent examples of such explanatory, causal models. These models show whole structures with different parts that interact and change over time and that are impacted by various forces or other structures, both internal and external. Psychologists and other social scientists should take as their goal the creation of such models, but they must recognize that reaching this end will require a richer definition of causality than the one supplied by positivism. In observation oriented modeling, Aristotle's four species of cause (material, efficient, formal, and final) in fact provide this more complete picture of causality. Final cause, in particular, proves important in subsequent chapters because it provides the means for modeling purpose, which is one of humanity's more enigmatic and important powers.

Finally, the realist claims that things have properties, some of which are essential for making a given thing to be what it is, whereas others are nonessential and referred to as accidental. By stating "this red book is 100 pages in length," for instance, a person is expressing a species concept, "book," that denotes the object's essence. Two accidental qualities of this particular book are that it is "red" and composed of "100 pages." Other books will naturally have different colors and numbers of pages, but their unity will still be recognized through the species concept of "book." For the Thomistic psychologists of the early 20th century, the starting point of inquiry was the essential unity of each person, and from this vantage point psychology was defined generically as a "study of the acts, powers, habits, and nature of man" (Brennan, 1941, p. 53). In other words, a person is a unified whole and the object of study; but as an organized composite of distinguishable parts, any given person must be understood using a variety of concepts, such as acts, powers, and habits. The aspects of personhood to which these concepts apply might also be referred to as qualities (or attributes), and from the realist perspective, any scientific investigation of these qualities must attend to their different forms of existence. Aristotle spoke of

a variety of different types of qualities, only some of which are quantitative; however, contemporary psychology is largely built on the premise that qualities such as intelligence, emotional states, and personality traits exist as continuous quantities.[5] This assumption will be challenged in the pages that follow, and it is not one made in observation oriented modeling.

Each of the aspects of philosophical realism discussed previously were largely foreign to the scholars who created the majority of research methods and statistical procedures used in modern psychology. Francis Galton, Karl Pearson, and Ronald Fisher, for instance, subscribed to various forms of positivism, which essentially denies that a person can know the things of nature. Drawing on the works of David Hume, John Stuart Mill, and the British empiricists, the starting point for developing a science of nature was considered to be the inner world of sense impressions. With such a science, one does not speak of the "things themselves" but, rather, sensations or impressions that are the objects of knowledge that can be labeled, correlated, ordered in time, and sorted into abstract laws. If regularities are noted in the various associated impressions, then predictions can be made about future events. Science is therefore best characterized as propositional—"if event A occurs, then event B will follow"—and statistics becomes the method of choice for modeling such linkages, which can never be held with certitude.

Against this backdrop, observation oriented modeling is introduced as a distinct alternative based on an entirely different philosophical position and approach toward data conceptualization and analysis. The Observation Oriented Modeling software and accompanying techniques that are also introduced in this book are therefore not to be construed as simply providing another set of tools that can be used to complement existing statistical methods. Rather, observation oriented modeling should be considered as fundamentally incompatible with the prevailing research tradition in psychology, and adopting this novel approach will require the abandonment of concepts that have been taught to students of psychology for more than 70 years. As argued in this book, such a change is necessary if psychology is to truly confront the limitations of its research tradition and subsequently set its foot firmly on a path toward a more genuinely scientific knowledge of persons.

[5] Aristotle offers a condensed list of qualities in his *Organon*, Categories, Chapter 8.

This page intentionally left blank

CHAPTER 2

Data at Its Core

Contents

Introduction 9
Deep Structure 10
Multiple Observations 20
Deep Structure Addition and Subtraction 22
Deep Structure Multiplication and Division 24
Logical Operations and Deep Structure 25

INTRODUCTION

At the heart of the observation oriented methods introduced in this book is the "deep structure" of all quantitatively and qualitatively ordered observations.[1] It is argued herein, and demonstrated in subsequent chapters with examples drawn from psychological research, that resolving ordered observations into their deep structure will not only create a gateway to simplified forms of data analysis but also lead to new insights into how data are to be collected and interpreted. With regard to simplification, establishing the common structure of all observations will permit the creation of a single class of data analytic methods that can be applied to a wide variety of theoretical questions and to ordered observations (i.e., "variables" using contemporary jargon) of virtually any type. Moreover, distinctions that have long been part of the terrain of social science measurement and statistical analysis, such as S. S. Steven's four scales of "measurement" (nominal, ordinal, interval, and ratio) and the different types of inferential errors (Type I, Type II, and Type III), will be made superfluous (Stevens, 1951, 1959).[2] With regard to gaining new insights,

[1] "Ordered" when paired with "observation" throughout this book indicates whole–part relations utilized by scientists when making systematic observations and when measuring. It is not meant to necessarily indicate judgments of quantity or of ordinal relationships. "Ordered observation" is similar to, but not equivalent to, the concept "variable," which is avoided here because its modern connotation points directly to Fisherian methods of analyzing variances.

[2] For a cogent and clear description of hypothesis testing and the different types of inferential errors, see Harris (1997).

Observation Oriented Modeling
ISBN 978-0-12-385194-9, Doi:10.1016/B978-0-12-385194-9.10002-7
© 2011 Elsevier Inc.
All rights reserved.

working with deep structures will reveal how contemporary modeling and analysis techniques, such as analysis of variance, multiple regression, and null hypothesis significance testing, are often ill-suited for the theoretical questions and empirical observations found in psychological, sociological, and educational journals. With Jacob Cohen and David Bakan (see Chapter 1), the argument could even be made on the basis of these insights that contemporary approaches work against the advancement of scientific knowledge.

The deep structure of ordered observations introduced in this chapter is therefore only part of a larger framework for thinking about empirical research and for modeling observations. So that this larger context does not get lost in mathematical details, a number of philosophical considerations are first presented in this chapter. The notion of deep structure is then presented via a simple example, and a variety of mathematical and logical operations appropriate for deep structures are introduced.

DEEP STRUCTURE

Western science can be traced from modernity back through time to the works of Anaximander, Democritus, Plato, Aristotle, and other ancient Greek thinkers. Pausing approximately midway in this historical journey at the Middle Ages, the first universities can be found springing from the cathedral schools in Europe,[3] and such men as Duns Scotus, Albert Magnus, Thomas Aquinas, and William of Ockham can be found breathing a new spirit into the many sciences—theological, philosophical, moral, political, and natural—of the day. The typical modern mind may have difficulty imagining a "science of theology," let alone a "science of morality," but the general use of the term *scientia* in medieval times conveyed a body of knowledge that was systematic and demonstrable via syllogism. In more specific uses, it was intended to mean certain knowledge of nature through its causes. Theological science was considered the "highest science" because it dealt with the first creating cause of the universe as well as the final end (i.e., the "final cause") in which happiness is attained, namely God.

[3] Interestingly, *universitas* was a legal term that originally referred to the whole body of students and teachers, not simply to the collection of sciences, or different topics of study. This focus on the persons in the first universities fits well with the methods introduced in this book, which are often centered round persons rather than variables or computed aggregates (Pace, 1912).

The modern academic can therefore understand how Aquinas arrived at the following conclusion:

> *Dicendum sacram doctrinam scientiam esse … quia procedit ex principiis notis lumine superioris scientiae, quae scilicet est Dei et beatorum.*[4]

An accurate understanding of how the medieval thinkers played such a critical role in the development of modern natural science can also be achieved. As argued persuasively by the historian of science, Stanley Jaki, various cultures produced "stillbirths" of science, whereas the Western mind alone, equipped with the appropriate theological and philosophical foundation laid by centuries of thought, produced the full flowering of natural science seen in the works of Gilbert, Galileo, Harvey, Newton, Lavoisier, and others (Jaki, 1978). The Schoolmen presupposed a deep unity among the sciences, and they could deftly cross-pollinate from one to the other without compromising their distinctiveness. With regard to natural science, they utilized philosophically and theologically informed notions of causality, uniformity, and motion that created a propitious intellectual framework for its development. Their thinking was also informed by notions of time, infinity, motion, and place that were the product of centuries of continuous thought from Aristotle to the Middle Ages. The gradual and sure progress of natural science from ancient Greece to medieval Europe, even through the so-called Dark Ages, has been carefully catalogued by Pierre Duhem, most notably in his *Système du monde*.[5]

For the current purposes, it is also important to note that many of the medieval Schoolmen relied heavily on Aristotle in their understanding of *scientia* as demonstration via syllogism and for their understanding of *causa*.[6] Specifically, cause was understood from a fourfold perspective: efficient, material, formal, and final. The paper on which these letters are printed (assuming the reader is not now gazing at a computer screen) can be

[4] "Sacred doctrine is a science … because it proceeds from principles established by the light of a higher science, namely, the science of God and the blessed" (St. Thomas Aquinas, 2008, Ia, 1, 2). All quotes from the *Summa* are taken from the Fathers of the English Dominican Province's translation, which can be retrieved online at http://www.newadvent.org/summa/index.html.

[5] See Duhem (1985) for English translations of selections from Duhem's *Le Système du monde*. It bears emphasizing that Duhem's work stands in stark contrast to the notion that science progresses through violent revolutions.

[6] William Wallace's careful scholarship clearly shows the Aristotelian nature of medieval science. See his *Prelude to Galileo* (1981), *The Scientific Methodology of Theodoric of Freiberg* (1959), and *Causality and Scientific Explanation*, *Volume I* (1972).

considered from this perspective by first noting that the paper is composed primarily of wood fibers, its material cause. Material reality is not formless, and the paper has a particular shape or form critical to what it is. This formal cause of the paper is one way it might be differentiated from other things made from the same material, such as cardboard or plywood. Efficient cause entails the movements from potentiality to actuality resulting in the paper—in other words, the sequence of steps taken through time to produce the paper (e.g., cutting down the trees that are potentially paper, chipping and pulping the wood, treating the pulp, and rolling the fibers into sheets). Lastly, the final cause of the paper is the purpose for which it was intended, in this case to serve as a medium for communication. Of course, Aristotle realized that most things in nature are not produced by people. He therefore also considered final cause to explain the end states normally attained by natural processes. For instance, trees grow to typical heights according to their species, and a red-hot piece of iron will return to its normal state if removed from a fire. Such natural end states and the other three causes are all needed to illuminate the fundamental nature of paper, trees, iron, and anything else that might be understood scientifically.[7]

Contemporary psychologists, however, typically overlook or ignore Aristotle's four causes. They instead rely on a restricted notion of cause codified in J. S. Mill's canons of logic (agreement, residue, difference, and concomitant variation), which were created to elucidate only efficient causes (see Chapter 9). Joseph Rychlak has spent most of his career in psychology tracing the history of causality and the limitations of relying only on efficient causes in modern theorizing and empirical research (Rychlak, 1985, 1988). The primary point of his argument is that a return to Aristotle's four causes could not only enlighten our psychological theories but also free them from the shackles of methods that are ill-suited for testing their central hypotheses. To demonstrate his point, over the course of his career Rychlak developed and successfully tested a theory of learning based largely on formal and final notions of causality (i.e., logical learning theory). Although he supported his theory with numerous experimental tests using methods equivalent to those utilized by psychologists considered more mainstream, the response from colleagues has largely been silence.

Despite this general indifference, Rychlak's historical and philosophical work has nonetheless shown that psychologists, social scientists, and natural

[7] Aristotle introduces and discusses the four causes in Book II, Chapter 3 of his *Physics*.

scientists have employed formal causes in their theories and research programs in ways that are not always made explicit. Two examples are considered here to support his arguments. First, the ascendancy of evolutionary psychology during the past 30 years has been driven almost exclusively by the notion that general predispositions can be passed down from one generation to the next via genes. In studies of twins, for example, researchers report greater similarity between identical twins than between fraternal twins in their responses to personality questionnaires and in their performances on tests of cognitive ability (Plomin, DeFries, McClearn, & McGuffin, 2008). But what is a gene? According to the official "Guidelines for Human Gene Nomenclature," it is "a DNA [deoxyribonucleic acid] segment that contributes to phenotype/function," and a DNA segment is an ordered sequence of four chemical bases: adenine, thymine, cytosine, and guanine (Wain et al., 2002). The ordered sequence itself is part of the larger DNA molecule that is in the shape of a double helix. If the sequence of the four bases is changed, or their location on the DNA molecule is changed, their contribution to phenotype or function will also change. Clearly, the emphasis here is on structures or patterns, and modern evolutionary psychology therefore relies on formal cause as much as material cause. Second, one of the most oft-told stories of imagination and creativity in science is that of August Kekulè's daydream of a snake seizing its own tail. If Kekulè's telling of this story is accepted as true, his dream played an important role in the discovery of the molecular structure of the benzene ring, namely six carbon atoms connected in a symmetrical ring with alternating single and double bonds. It is worth noting that the empirical formula for benzene was known at the time; hence, it was specifically the structure (i.e., the formal cause of benzene) that Kekulè discovered that opened the gates for further developments in the science of aromatic compounds.

Scientific progress, then, is more than developing ways to explain quantitative change over time (a question of efficient cause); it is also marked by increases in knowledge regarding the form of matter. On an even deeper level, the process of quantitative measurement can be viewed metaphysically to presuppose matter and form and to emerge from an application of the "one–many" dialectic that intrigued Greek and Medieval thinkers.[8] These metaphysics are briefly explored in Chapter 8, but here it is sufficient to

[8] The metaphysical basis of measurement as an application of the "one and many" dialectic by Aristotle and Thomas Aquinas was advanced by Charles Crowley (1996).

point out that formal cause is not an esoteric, outdated concept that is completely foreign to the workings of the modern scientific mind. Indeed, it is at least tacitly employed in the theories and methods of scientists, and it can stimulate ideas such as "pattern," "organization," and particularly "whole–part relations" that, as Rychlak has suggested, may shed new light on contemporary techniques of measurement and analysis.

Beginning simply, the so-called Likert-type item, which has been a staple of the social sciences for more than 70 years, can be examined from the vantage point of the "one–many" dialectic or from the perspective of whole–part relations (Likert, 1932). The following example item might be found on any number of self-report questionnaires:

Generally speaking, I like to spend most of my time alone.

Highly Disagree	*Disagree*	*Neither Disagree nor Agree*	*Agree*	*Highly Agree*
____	____	____	____	____

This item expresses an intention of its author (e.g., to learn something interesting about the nature of the human psyche), who intuitively recognizes the singular and plural nature of the item—that is, one item with many parts. This intuitive understanding of whole–part relations can be seen in the way the item is carefully demarcated from other text, and potentially other items, on the same page. Imagine if the item were presented with a second item as follows:

> *Generally speaking, I like to spend most of my time alone highly disagree disagree neither disagree nor agree agree highly agree I am the type of person who likes to see the good in every person highly disagree disagree neither disagree nor agree agree highly agree*

Obviously, the two items are nearly indiscernible and become lost in the text. Only by demarcating the items through careful formatting will their whole and separate identities be recognized. The particular parts of the previous item, namely its stem ("Generally speaking …") and anchored rating options, are also recognized in the way the item is formatted by its author. The stem is presented as the first part of the item to be attended to above the anchors; the anchors are separated so that "highly agree," for instance, is discernible from "agree"; and the rating option lines are aligned under the verbal anchors. With minimal training and experience, the person responding to the item is able to recognize it as a whole item, separate from other items and text that might be on the same page. The person will then

evaluate its parts to reach some state of analogous unity between a part of the item ("highly agree," "disagree," etc.) and a conscious intention, indicating the point of unity with a check mark on the corresponding line under the anchor. The mental processes involved in the conscious intention are never observed by the item's author, and it is always possible that the person's intention may be inconsonant with the intention of the author. For example, the person may respond to the item intending to deceive the author of the item or may select a point on the scale as a result of some other intention, such as selecting "highly disagree" to avoid a current state of sadness associated with acknowledging a true desire to be alone. These considerations are treated as threats to validity in modern psychometric theory and are only mentioned here to be clear about what is observed and what is not observed when a person quietly places a check mark on the scale.

The previous Likert-type item can be conventionally "scored" by assigning integers to the scale points. For instance, integers ranging from -2 to +2 can be assigned, such that a response of "highly disagree" is scored -2, whereas a response of "highly agree" is scored +2. The range of the values is arbitrary as long as the points on the scale are separated by a constant difference, usually 1. Values of 0 to 5 or 1 to 6 would therefore also be common for this type of item, with high values indicating greater subjective agreement with the statement found in the item stem. High values could furthermore be interpreted as indicative of "introversion," whereas low values would be indicative of "extraversion."

In light of the whole–part reasoning discussed previously, a different way to score the item is suggested that recognizes its whole structure and the person's particular response. Consider a person, for example, who endorses "disagree" on the previous item. All five options of the item can be coded individually as "0" or "1," with the latter indicating the option on the scale selected by the person; that is,

0 1 0 0 0.

The whole item is thus coded as a string of values representing its individual parts, which are in turn coded to indicate the unitary part judged by the person as representing his or her intention. A person endorsing "highly disagree" would be coded as

1 0 0 0 0,

a person responding "neither disagree nor agree" would be coded as

0 0 1 0 0,

and so forth. This novel representation or coding of the item is referred to as its "deep structure," and each individual part of the whole is referred to as a "unit." Obviously, the previous item is composed of five units.

As another example, Table 2.1 shows how the deep structure of biological sex can be represented ("1 0" for males and "0 1" for females). The original categories for biological sex do not reveal any differences in quantity (i.e., "male" is not greater than or less than "female"), and consequently the assignment of values in the deep structure is arbitrary. In other words, males could be represented as "0 1" and females as "1 0."

The arbitrary nature of coding deep structure for nonquantitatively ordered observations can also be seen in Tables 2.2 and 2.3 for ethnicity. Either deep structure would suffice for representing the ethnic groups delineated by the categories. The only constraint is that each category of ethnicity must be uniquely identified by a deep structure string of values composed of a single 1 and multiple 0's. It can be seen in Tables 2.2 and 2.3 that no two ethnic groups are coded alike. Deep structures for ordered observations that are assumed to indicate ordered quantities, like the Likert-type item discussed previously, or equal interval quantities will also uniquely identify each point on the measurement scale. In such instances, however, the coding for the scale points will not be arbitrary if mathematical operations performed on deep structures are to be considered legitimate. With quantitatively ordered observations, the ordinal position of the "1" in each string of values must correspond to the unit's ordinal position within the whole. The deep structure for the previous Likert-type item, for example, which is assumed to convey some aspect of quantity (i.e., a person selecting "highly agree" is considered more introverted than a person selecting "highly disagree"), is reported in Table 2.4. If the quantitative structure of the ordered observations is to be preserved, no other method of coding the deep structure is permissible.

The deep structure of ordered observations that more clearly indicate continuous quantity, such as the standard measures of length, weight, and

Table 2.1 Deep structure for observations ordered as "male" and "female"

Biological sex	Deep structure	
Male	1	0
Female	0	1

Table 2.2 Deep structure for observations ordered as different ethnic groups

Ethnicity	Deep structure					
African American	1	0	0	0	0	0
Asian	0	1	0	0	0	0
Caucasian	0	0	1	0	0	0
Hispanic	0	0	0	1	0	0
Native American	0	0	0	0	1	0
Other	0	0	0	0	0	1

Table 2.3 Alternative deep structure for observations ordered as different ethnic groups

Ethnicity	Deep structure					
African American	0	1	0	0	0	0
Asian	0	0	1	0	0	0
Caucasian	1	0	0	0	0	0
Hispanic	0	0	0	0	0	1
Native American	0	0	0	0	1	0
Other	0	0	0	1	0	0

Table 2.4 Deep structure for a 5-point Likert-type item

Response	Deep structure				
(−2) Highly disagree	1	0	0	0	0
(−1) Disagree	0	1	0	0	0
(0) Neither	0	0	1	0	0
(+1) Agree	0	0	0	1	0
(+2) Highly agree	0	0	0	0	1

speed, can also be determined. In order to demonstrate how the deep structure of such measures can be coded legitimately, consider a child who is asked to measure the following line in millimeters using a ruler that is 20 cm in total length:

Although it may be taken for granted, implicit in the request is the assumption that the girl will perceive the line as a distinguishable whole, with two terminal points, that can be divided in her mind into parts. Nonetheless, it is easy to imagine the child laying the ruler on the page and then scanning for the point on the ruler's scale marked "0." The whole ruler

comes into play as a unified "measuring device" in this act, as does the whole scale demarcated on its surface. In other words, through previous experience she understands that the ruler can be used for such tasks and that its scale is demarcated into equal segments (or parts) that on one end indicate low quantities and on the other end indicate high quantities. She begins by aligning, as closely as possible, the segment marked "0" on the ruler with the left side of the line. She then scans the ruler for the point that matches the end of the line. When she approaches the end of the line, she may shift her judgment back and forth between two marked points on the ruler until she arrives at a judgment of "oneness" between a particular mark and the end of the line. Her final judgment of length thus implicitly entails the whole ruler and the exclusion of all points on the scale (judged as not "one" with the end of the line) except the point she selects. The deep structure of her response would thus entail 201 units: 200 zeros, including the zero point on the scale, and a single "1" corresponding to the point on the scale she has selected. Admittedly, the deep structure of such measurements can become greatly complex with precise scales, but this pragmatic detail does not contradict the principle of converting continuously structured quantitative observations to deep structure form.

With this rudimentary understanding of deep structure in place, several points deserve attention. First, when speaking of the principles that underlie continuous quantitative measurement, such as equality, uniformity, and regularity, it is assumed that these principles are appropriate to the matter under study. For instance, the previously mentioned girl relied on these principles when measuring the line because they were "built into" the ruler's scale; that is, the whole scale is equal to the sum of its parts, which are equal to one another, and the scale lies on a perfectly straight (uniform) line. Obviously, when classifying persons as "male" or "female" or sorting them into different groups based on their ethnic heritage, the same set of principles may not necessarily hold. Critics of contemporary psychometric models have even gone so far as to suggest that the principles of measurement traditionally employed by scientists (particularly physicists) are not applicable to the vast majority of ordered observations made in psychology and the other social sciences.[9] Realizing that all such observations can be expressed in a common deep structure form does not in any way answer these critics' valid charges, but it may provide new pathways

[9] Joel Michell (1999) has made the most convincing arguments regarding the untested or unmet quantitative assumptions underlying statistical models in the social sciences.

toward thinking about psychological measurement in the future. The topic of measurement and these controversies are treated more explicitly in Chapter 8.

Second, the current methods may not be very useful for scientists working with truly continuous quantitatively ordered observations based on large-scale ranges and precise and reliable measurement. The issue is not the potentially unwieldy nature of the deep structure of such measurements, such as the 20-cm ruler presented previously, but rather the simple fit between theory and method. The observation oriented methods introduced in this book—based on the deep structures of ordered observations—are simply more likely to provide clearer tests of the theoretical and hypothetical statements of social scientists than of physical scientists working with qualities that are structured as continuous quantities. It is also argued in subsequent chapters that the current methods are superior to Karl Pearson's correlational methods and Ronald Fisher's inferential variance-based methods employed by most psychologists and social scientists today.

Lastly, recognition of the deep structure underlying ordered observations leads to a finessed understanding of reliability in social science measurement. According to the classical true score model that undergirds much of modern psychometric theory, reliability is defined generally as consistency of measurement and specifically as the ratio of true score to observed score variance in a set of quantified observations. As a proportion, an estimate of reliability will range from 0 to 1, with high values indicating greater reliability. Another model of reliability, referred to as generalizability theory, similarly employs notions of variability to arrive at estimates of reliability for a set of quantified observations. With deep structures and observation oriented models, the contemporary notions of reliability will be distinguished from precision of measurement. Imagine if the previously mentioned girl wished to make a more precise measurement of the line. She could proceed to make numerous observations with the same ruler and then calculate an average; however, this would concern the validity of her measurement. If she were to examine the variability of her multiple readings, this would tell her something of the consistency of her use of the scale, but it would still not increase her precision. Precision, as defined in any standard dictionary, refers to being minutely exact; therefore, the only way to increase her precision of measurement is to use a different ruler with a more exact scale for ordering the observations (e.g., a hypothetical ruler marked in nanometers). Understanding the distinction between consistency and precision permits social scientists to approach the analysis of their

ordered observations from a more flexible position. For example, a psychologist may estimate the test–retest reliability of her 100-point intelligence test to be rather low; however, she may also find that using a grouped, 10-point version of the same test yields ordered observations that are much more consistent. The psychologist can test such alternatives by simply defining the deep structure of the original test with 100 units or with only 10 units in which the original scale points are combined into groups (1–10, 11–20, etc.). A number of examples demonstrating this application of deep structures are presented in subsequent chapters.

MULTIPLE OBSERVATIONS

Inspired by Aristotle's notion of formal cause, deep structure thus provides a novel way of thinking about ordered observations or data. The full impact of this vantage point, however, cannot be realized until the deep structure of multiple observations is understood. A single data point of course never suffices in science; multiple observations are required for replication, to establish generalizability, and to ensure consistency and precision of measurement. For observation oriented modeling, data will be routinely arranged in matrix form with the observations as rows and deep structure units as columns. Table 2.5 shows the deep structure for nine people who reported their biological sex as "M" or "F," and Table 2.6 shows the deep structure for their responses to the extraversion/introversion Likert-type item presented previously. Using standard conventions for recording data in matrices, the two deep structures can be presented as follows:

$$\text{Gender} = {}_9\mathbf{Y}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad \text{Extraversion} = {}_9\mathbf{X}_5 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Table 2.5 Deep structure for the reported biological sex of nine persons

		Deep structure	
Person	**Biological sex**	**$Scale_M$**	**$Scale_F$**
P_1	M	1	0
P_2	F	0	1
P_3	M	1	0
P_4	M	1	0
P_5	F	0	1
P_6	F	0	1
P_7	M	1	0
P_8	F	0	1
P_9	M	1	0

Table 2.6 Deep structure for nine persons' responses to a 5-point Likert-type scale

		Deep structure				
Person	**Scored response**	**$Scale_{-2}$**	**$Scale_{-1}$**	**$Scale_0$**	**$Scale_{+1}$**	**$Scale_{+2}$**
P_1	−1	0	1	0	0	0
P_2	2	0	0	0	0	1
P_3	−2	1	0	0	0	0
P_4	−1	0	1	0	0	0
P_5	1	0	0	0	1	0
P_6	2	0	0	0	0	1
P_7	−2	1	0	0	0	0
P_8	1	0	0	0	1	0
P_9	−1	0	1	0	0	0

In this book, matrices are presented as bold capital letters, and subscripts are used to denote their dimensionality. For instance, the deep structure matrix for biological sex is referred to as ${}_9\mathbf{Y}_2$, a 9×2 matrix. Particular cells of a matrix are referenced with subscripted numbers in brackets following the matrix letter. For example, the first row and first column entry for ${}_9\mathbf{Y}_2$ is $\mathbf{Y}_{[11]} = 1$. $\mathbf{Y}_{[12]} = 0$, $\mathbf{Y}_{[21]} = 0$, $\mathbf{Y}_{[22]} = 1$, etc.

Once in matrix form, these deep structures can be manipulated using matrix algebra and other unique operations. A method particularly important for observation oriented modeling is a simplified form of Procrustes rotation. This method will provide the primary mechanism for modeling relationships between deep structures such as ${}_9\mathbf{Y}_2$ and ${}_9\mathbf{X}_5$ presented previously. Before introducing this technique in Chapter 3, however, a number of deep structure matrix operations are described.

DEEP STRUCTURE ADDITION AND SUBTRACTION

Simple addition can be emulated with deep structure matrices that are constructed from discrete or continuous quantities. Consider a scale that ranges in whole numbers from 1 to 4 and a second scale that ranges in whole numbers from −1 to 1. Scores on the scales and their sums and differences are as follows:

Person	First Scale	Second Scale	Sum	Difference
P_1	3	−1	2	4
P_2	4	1	5	3
P_3	3	0	3	3
P_4	1	−1	0	2
P_5	3	1	4	2

The deep structure of the responses on each measure follows:

	First Scale				Second Scale		
Person	$Scale_1$	$Scale_2$	$Scale_3$	$Scale_4$	$Scale_{-1}$	$Scale_0$	$Scale_1$
P_1	0	0	1	0	1	0	0
P_2	0	0	0	1	0	0	1
P_3	0	0	1	0	0	1	0
P_4	1	0	0	0	1	0	0
P_5	0	0	1	0	0	0	1

With normal addition, the possible range of values for the sums of the two scales is 0 to 5, or six different values. By comparison, the two deep structures form seven total units. Considered generally, the number of total units for two deep structures will exceed the number of possible values computed through normal addition by one. The first step to deep structure addition, then, is to generate a new deep structure matrix that has p (the number of persons, or observations) rows and $j + k - 1$ units, where j and k equal the number of points on the first and second scales, respectively. The second step is then to sum the ordinal positions of the 1's in the two original deep structure matrices. The sums, minus one, indicate the ordinal positions for recording 1's in the new deep structure matrix. For example, the 1's in the deep structure matrices for the first person are recorded in the third and

first columns for the first and second scales, respectively. Taking $3 + 1 - 1$ yields 3, and thus a 1 is recorded in the third column of the new deep structure, which corresponds to a value of 2. Similarly, the 1's in the deep structure matrices for the second person are recorded in the fourth and third columns, resulting in $4 + 3 - 1 = 6$. In the new deep structure, a 1 is recorded in the sixth column for this person. All of the results obtained from this deep structure addition follow:

	Summed deep structure					
Person	**$Scale_0$**	**$Scale_1$**	**$Scale_2$**	**$Scale_3$**	**$Scale_4$**	**$Scale_5$**
P_1	0	0	1	0	0	0
P_2	0	0	0	0	0	1
P_3	0	0	0	1	0	0
P_4	1	0	0	0	0	0
P_5	0	0	0	0	1	0

Subtraction works in similar fashion. The differences obtained from subtracting scores on the second scale from scores on the first scale are shown in the first table in this section. The possible range of these differences is 0 to 5, or five integer values and zero. When subtracting deep structures, the ordinal positions of the 1's are again determined for each person, and the new ordinal position is computed through subtraction. For example, 1's in the deep structure matrices for the third person are recorded in the third and second columns. These values are subtracted and added to the number of columns in the second deep structure, namely $3 - 2 + 3 = 4$. In the new deep structure, a 1 is thus recorded in the fourth column for this person, which corresponds to a value of 3. For the fourth person, the result is computed as $1 - 1 + 3 = 3$, the third column of the new deep structure (corresponding to a scale value of 2). The remaining results follow:

	Subtracted deep structure					
Person	**$Scale_0$**	**$Scale_1$**	**$Scale_2$**	**$Scale_3$**	**$Scale_4$**	**$Scale_5$**
P_1	0	0	0	0	1	0
P_2	0	0	0	1	0	0
P_3	0	0	0	1	0	0
P_4	0	0	1	0	0	0
P_5	0	0	1	0	0	0

Note that deep structure addition and subtraction for discrete and continuous quantity observations are associative and commutative.

As alluded to previously, and as discussed more fully in Chapter 8, psychologists routinely presume that Likert-type scales, rating scales, and scores on standardized tests represent qualities in nature (e.g., intelligence, depression, and personality traits) that are structured as continuous quantities. If they are not structured in such a manner, then basic mathematical operations such as addition, subtraction, or the computation of a mean are not legitimate for the observations. Of course, there will be nothing to stop the psychologist from plugging numbers into a computer or calculator and performing such operations, but the conclusions reached with regard to the qualities being studied will likely be erroneous. Accordingly, deep structure addition and subtraction can be performed on Likert-type scales and other types of ratings, but the resulting deep structure should be interpreted as one that combines and preserves the ordering of the units of observation. Even ordered units that combine observations can be "added" or "subtracted" in the previously discussed manner as long as the units are considered to be ordered from least to most in some way. For instance, a 100-point scale could be divided into units of 10 (1–10, 11–20, etc.), and the deep structure of these ordered observations could be "added" to other observations. "Added" is placed in quotation marks here to indicate that the operations would not represent true addition in this instance but, rather, a series of operations that combine and preserve the ordering of the units of observations. As demonstrated in subsequent chapters, psychologists often work with additive models; most notably in multiple regression. Similar types of models can be tested within the framework of observation oriented modeling using deep structure addition and subtraction as long as the units of observation are reasonably ordered from least to greatest.

DEEP STRUCTURE MULTIPLICATION AND DIVISION

A method for accomplishing multiplication or division through deep structures is not readily apparent. First, it is not clear how the number of columns for the new deep structure matrix could be determined without reference to the original scale values. Second, negative numbers seem to be particularly problematic. A solution might be found in group theory, Peano's theories, or the known relationships between addition and

multiplication. Fortunately, even without such a solution in hand, multiplicative models can be tested using observation oriented modeling techniques developed in this book, but the deep structure of the products (or quotients) cannot be generated automatically as with addition or subtraction. Instead, the multiplication or division must first be performed on the original scale values and then the deep structure determined for the resulting products or quotients. Obviously, as with addition and subtraction, such mathematical operations will require the psychologist to work with countable, discrete units or to demonstrate the continuous quantitative structure of the attribute or quality being measured.

LOGICAL OPERATIONS AND DEEP STRUCTURE

One of the intriguing features of deep structure is that logical operations can readily be performed on binary ordered observations. Table 2.7 shows the self-reported biological sex (ordered as male or female) and handedness (ordered as only left or right) for seven persons along with the deep structures for the responses.

Table 2.7 Original data and deep structure for seven persons' biological sex and handedness

			Deep structure		Deep structure	
Person	**Biological sex**	**Handedness**	**$Scale_M$**	**$Scale_F$**	**$Scale_L$**	**$Scale_R$**
P_1	M	Left	1	0	1	0
P_2	F	Right	0	1	0	1
P_3	M	Left	1	0	1	0
P_4	M	Right	1	0	0	1
P_5	F	Right	0	1	0	1
P_6	F	Left	0	1	1	0
P_7	M	Right	1	0	0	1

Logical combinations of biological sex and handedness can be expressed in deep structure form. For instance, the conjunctive statement "Male *and* Left-Handed," or generically "A *and* B," can be represented in deep structure form by creating a new truth value deep structure matrix. In the new matrix, "1" indicates the logical statement is true, whereas "0" indicates the statement is false. The conjunctive deep structure matrix, where $\wedge$ indicates "and," follows:

	Deep structure	
Person	**M ∧ L**	**F ∨ R**
P_1	1	0
P_2	0	1
P_3	1	0
P_4	0	1
P_5	0	1
P_6	0	1
P_7	0	1

As can be seen, the first and third persons are both male and left-handed, whereas everyone else is either female or right-handed, where ∨ indicates "or". The same general strategy can be employed to examine the deep structure of conjunctive statements between any pair of binary ordered observations.

Now consider the disjunctive statement "Male *or* Right-Handed" or generically "A *or* B." Again, a new truth value deep structure matrix is created in which 1's indicate the disjunctive statement is true for the persons and 0's indicate the statement is false. The deep structure for the statement "Male or Right-Handed" for the seven persons in Table 2.7 follows:

	Deep structure	
Person	**M ∨ R**	**F ∧ L**
P_1	1	0
P_2	1	0
P_3	1	0
P_4	1	0
P_5	1	0
P_6	0	1
P_7	1	0

Only the sixth person is a left-handed female.

A biconditional (A ↔ B) statement is often described as "A if and only if B" or "A iff B." If A and B have the same truth value (i.e., the same deep structure values), then the biconditional statement is true. If they have opposite truth values, then the statement is false. The deep structure for the biconditional statement "Female iff Right-Handed" for the observations in Table 2.7 is therefore as follows:

	Deep structure	
Person	**F iff R**	**~(F iff R)**
P_1	1	0
P_2	1	0
P_3	1	0
P_4	0	1
P_5	1	0
P_6	0	1
P_7	0	1

The deep structures of conditional ($A \rightarrow B$) statements can also be determined. These statements are read as "If A then B," and their truth value is equivalent to "$\sim A \vee B$." For instance, "If Male then Right-Handed" is equivalent to "*not* Male *or* Right-Handed," or "Female *or* Right-Handed." The deep structure is therefore as follows:

	Deep structure	
Person	**M → L**	**~(M → L)**
P_1	0	1
P_2	1	0
P_3	0	1
P_4	1	0
P_5	1	0
P_6	1	0
P_7	1	0

Finally, exclusive disjunction statements are of the form "A *or* B, but *not* both," such as "This person is female *or* right-handed, but *not* both." The symbol XOR is often used to denote exclusive disjunction. For "Female XOR Right-Handed," then, the deep structure is as follows:

	Deep structure	
Person	**F XOR R**	**~(F XOR R)**
P_1	0	1
P_2	0	1
P_3	0	1
P_4	1	0
P_5	0	1
P_6	1	0
P_7	1	0

When psychologists are not working with discrete countable units or continuously structured quantities, such logical statements will likely come to the forefront. They can certainly be seen in the medical model of classifying psychological disorders according to the presence or absence of different symptoms; and observation oriented modeling offers a flexible methodology for evaluating the empirical support for such models, examples of which are presented in subsequent chapters.

CHAPTER 3

Rotating Deep Structures

Contents

Introduction 29
Procrustes Rotation 29
Binary Procrustes Rotation 36

INTRODUCTION

In the previous chapter, the common deep structure that underlies all ordered observations, both quantitative and nonquantitative, was presented. Different mathematical and logical operations that could be applied to the deep structure of different sets of observations were also discussed. In order to finally breathe life, so to speak, into these deep structures in Chapter 4, the primary mathematical technique underlying observation oriented modeling must first be introduced, namely a simplified form of Procrustes rotation referred to as binary Procrustes rotation. As will be shown, this technique can be used to express one deep structure in the form of another. Considering the last example from Chapter 2, for instance, the goal might be to rotate the deep structure of handedness into conformity with the deep structure of biological sex. In Chapter 4, binary Procrustes rotation will in fact be used to perform such transformations in the context of testing research hypotheses put forth by psychologists. The result will be a new approach for modeling that places focus on the observations as they are ordered in the study rather than on numerical aggregates and variables.

PROCRUSTES ROTATION

In the realms of factor analysis and general linear algebra, the term "rotation" refers to the rigid transformation of a set of vectors. This definition can be understood fully via an example of two people who rate themselves on the following scales:

Withdrawn ___ ___ ___ ___ ___ Outgoing
–2 –1 0 1 2

Gregarious ___ ___ ___ ___ ___ Shy
–2 –1 0 1 2

Observation Oriented Modeling
ISBN 978-0-12-385194-9, Doi:10.1016/B978-0-12-385194-9.10003-9
© 2011 Elsevier Inc.
All rights reserved.

The values below the scales indicate the scores for each point, and the first item is posited as an indicator of extraversion (i.e., values of 1 and 2 indicate extraversion), whereas the second is considered to be an indicator of introversion. The first person scores 2 and -2 on the items, respectively, thus indicating extraversion; the second person scores -1 and 2 on the two items, respectively, thus indicating introversion. The persons' responses can be recorded in a 2 × 2 matrix and then plotted in a standard Cartesian coordinate space, both shown in Figure 3.1. The points are connected to the origin of the graph by arrows, which are referred to as vectors in algebraic jargon. The vectors can be rotated by multiplying the matrix of scores, ${}_2\mathbf{Y}_2$, by what is known as a transformation matrix, $\mathbf{T}$. Given the transformation matrix

$$ {}_2\mathbf{T}_2 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, $$

for instance, the values of ${}_2\mathbf{Y}_2$ are transformed to

$$ {}_2\mathbf{X}_2 = {}_2\mathbf{Y}_2\,{}_2\mathbf{T}_2 = $$
$$ {}_2\mathbf{X}_2 = \begin{bmatrix} 2 & -2 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ -1 & -2 \end{bmatrix}. $$

The values in ${}_2\mathbf{X}_2$ are plotted in Figure 3.2, which reveals that this particular transformation rotated the vectors counterclockwise 90°. It is also important to note that the angle between the two vectors, which is 160°, did

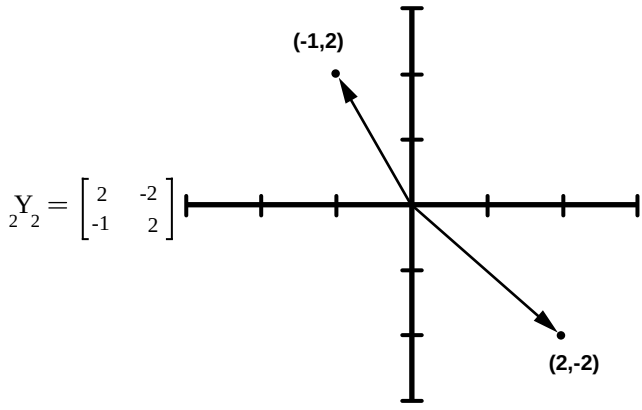

Figure 3.1 Matrix and two-dimensional plot of responses to two questionnaire items

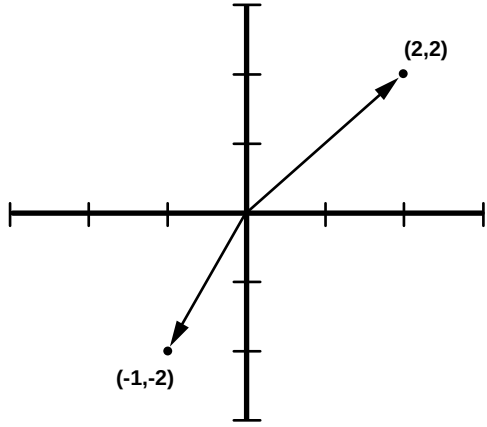

Figure 3.2 Rotated vectors from Figure 3.1

not change as a result of the transformation, nor did the lengths of the vectors. This consistency explains why the rotation was previously referred to as "rigid." Any number of similar rotations could be accomplished by simply altering the values in the transformation matrix.

Rotation can also be thought of from the perspective of the axes forming the two-dimensional space. For the data in Figure 3.1 and the transformation matrix presented previously, the 90° counterclockwise rotation of the vectors can be considered as a 90° clockwise rotation of the axes. The resulting graph would be the same as Figure 3.2, but high scores on the *x*-axis and high scores on the *y*-axis would both indicate extraversion, whereas low scores on both axes would indicate introversion. In short, the rotation in this example accomplished the goal of reverse keying the second item.

A specific type of transformation encountered in psychology, particularly in personality psychology, is Procrustes rotation. The rotation shown previously transformed a single set of vectors. With Procrustes rotation, the goal is to rotate two sets of vectors to maximal agreement in a common space. As an example, consider the following ratings from Ben and Erika regarding three brands of breakfast cereal and a subjective "ideal cereal":

		Cereal A	Cereal B	Cereal C	Ideal cereal
Ben	Sweet	−3	−2	2	3
	Crunchy	1	2	2	2
Erika	Expensive	−2	−3	−2	−1
	Healthy	−2	−1	2	2

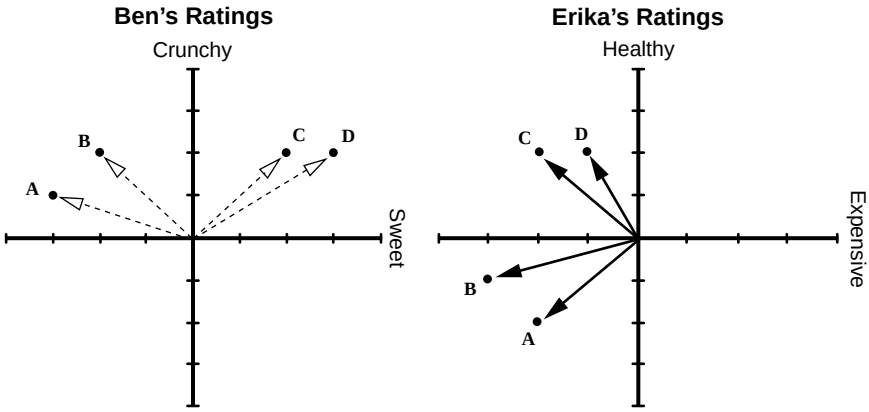

Figure 3.3 Two-dimensional plots of rated cereals on 5-point scales

These ratings can be plotted in two different two-dimensional spaces, as shown in Figure 3.3. It can readily be seen that Ben and Erika agree with respect to the ways they group the cereals into clusters: Cereals A and B are seen as similar to one another, and cereals C and D are seen as similar to one another. It can also readily be seen that if the axes for Ben's space are rotated 90° counterclockwise, the vectors for his ratings will match very closely with Erika's. The goal of Procrustes rotation is, in fact, to rotate these two sets of vectors (or axes) to maximal agreement as determined by some criterion, such as least squares.

In his text on factor analysis, Stanley Mulaik credits Green with first publishing a method for deriving the transformation matrix, **T**, necessary for rotating one set of vectors to another. He also credits Hurley and Cattell with popularizing the name "Procrustes" for this specific type of transformation and discusses important papers by Schönemann from the 1960s that generalized Green's early work (see Mulaik, 1972, pp. 293–299; see also Green, 1952; Hurley & Cattell, 1962; Schönemann, 1966). These generalizations serve as the basis for most modern applications of Procrustes rotation and for the current presentation.

The transformation begins by designating one matrix of values as the target matrix, here defined as ${}_p\mathbf{Y}_v$ (where p indicates the number of rows in the matrix, and v indicates the number of columns), and the second matrix as the conforming matrix, here defined as ${}_p\mathbf{X}_m$.[1] The conforming matrix is

[1] The terminology "conforming matrix" is not generally used in discussions of Procrustes rotation. Our choice of the term will be consistent with the way this particular transformation will be used in modeling relationships between ordered observations.

rotated to the target matrix by first performing a singular-value decomposition on the product of ${}_m\mathbf{X}'_p$ and ${}_p\mathbf{Y}_v$,

$$ {}_m\mathbf{X}'_p \, {}_p\mathbf{Y}_v = {}_m\mathbf{U}_v \, {}_v\mathbf{L}_v \, {}_v\mathbf{V}'_v, \quad [3.1] $$

where ${}_m\mathbf{U}_v$ and ${}_v\mathbf{V}_v$ are column-orthonormalized eigenvectors, and ${}_v\mathbf{L}_v$ is a diagonal matrix of singular values (similar to eigenvalues). The values in these three matrices are computed from one of several techniques, such as Jacobi's method or the QR method, which need not be discussed here.[2] A transformation matrix is then computed:

$$ {}_m\mathbf{T}_v = {}_m\mathbf{U}_v \, {}_v\mathbf{V}'_v. \quad [3.2] $$

Finally, the conforming matrix is multiplied to the transformation matrix to compute the rotated values, ${}_p\mathbf{B}_v$,

$$ {}_p\mathbf{B}_v = {}_p\mathbf{X}_m \, {}_m\mathbf{T}_v. \quad [3.3] $$

Treating Ben's ratings as the conforming matrix and Erika's as the target matrix, the transformed values are obtained from the Procrustes rotation,

$$ {}_2\mathbf{B}_4 = \begin{bmatrix} -1.77 & -2.47 & -1.38 & -1.11 \\ -2.62 & -1.38 & 2.47 & 3.43 \end{bmatrix}, $$

which are highly similar to Erika's ratings:

$$ {}_2\mathbf{Y}_4 = \begin{bmatrix} -2 & -3 & -2 & -1 \\ -2 & -1 & 2 & 2 \end{bmatrix}. $$

Figure 3.4 shows the high correspondence between Erika's original and Ben's rotated ratings.

The previous equations are perfectly general in the sense that Erika and Ben could have rated many more cereals on many more dimensions. What is apparently not widely known, however, is that the transformation can even be used if they had rated the same cereals on different numbers of dimensions. Ben, for instance, could have rated the cereals on three different dimensions instead of two, thus producing a 3×4 matrix of ratings. Erika's ratings would

[2] Any standard textbook of linear algebra will likely provide discussions of the different methods for computing eigenvectors and eigenvalues from a matrix. Mulaik (1972) also provides a summary of the different methods.

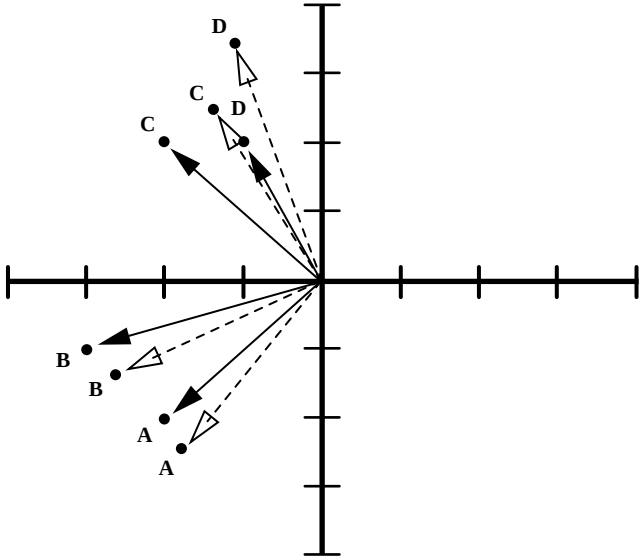

Figure 3.4 Plot of Erika's ratings (solid vectors) and Ben's rotated vectors (dashed vectors)

still be recorded in a 2×4 matrix, and because their two matrices are matched with regard to the number of columns (four cereals), they can still be rotated to maximal agreement. Such a transformation is tantamount to expressing a three-dimensional space (Ben's) in a two-dimensional space (Erika's). The order could be reversed as well so that Erika's matrix is rotated to maximal agreement with Ben's. In this case, a two-dimensional space would be rotated to a three-dimensional space. It should be clear that the two rotations, Ben to Erika and Erika to Ben, are not necessarily associative when the dimensionality of the matrices does not match; in other words, they could produce results that are quite different. Nonetheless, Procrustes rotation is not restricted to equal dimensionality between pairs of matrices; the two matrices need only be matched on one dimension.

This aspect of Procrustes rotation will play an important role in the observation oriented models to be introduced later, and it is worth demonstrating with another example. Specifically, consider the following two matrices:

$$ {}_2\mathbf{Y}_2 = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \; {}_2\mathbf{X}_3 = \begin{bmatrix} -.707 & -1 & .707 \\ .707 & -1 & -.707 \end{bmatrix} $$

${}_2\mathbf{Y}_2$ is the target matrix, ${}_2\mathbf{X}_3$ is the conforming matrix, and their vectors are shown in Figure 3.5.

Procrustes rotation can be used to rotate the vectors to maximal agreement even though the first set of vectors is plotted in two dimensions and the second set is plotted in three dimensions. Employing Eq. [3.1] results in the transformation matrix,

$$ {}_3\mathbf{T}_2 = \begin{bmatrix} .707 & 0 \\ 0 & -1 \\ -.707 & 0 \end{bmatrix}, $$

which is then multiplied to $\mathbf{X}$:

$$ {}_2\mathbf{X}_3\ {}_3\mathbf{T}_2 = {}_2\mathbf{B}_2 = \begin{bmatrix} -.707 & -1 & .707 \\ .707 & -1 & -.707 \end{bmatrix} \begin{bmatrix} .707 & 0 \\ 0 & -1 \\ -.707 & 0 \end{bmatrix} $$

$$ = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}. $$

The result is that ${}_2\mathbf{B}_2$ equals ${}_2\mathbf{Y}_2$, and the conforming matrix was in this case rotated to perfect agreement with the target matrix. Examining the individual entries in the ${}_2\mathbf{X}_3$ and ${}_3\mathbf{T}_2$ matrices that yield the products in ${}_2\mathbf{B}_2$ reveals that the first and third columns in ${}_2\mathbf{X}_3$ are combined while the second column is reflected, resulting in the same values as ${}_2\mathbf{Y}_2$ plotted on the left side of Figure 3.5. Conversely, ${}_2\mathbf{Y}_2$ can be rotated to ${}_2\mathbf{X}_3$, and in this instance, perfect agreement will again be obtained as ${}_2\mathbf{B}_3$ equals ${}_2\mathbf{X}_3$.

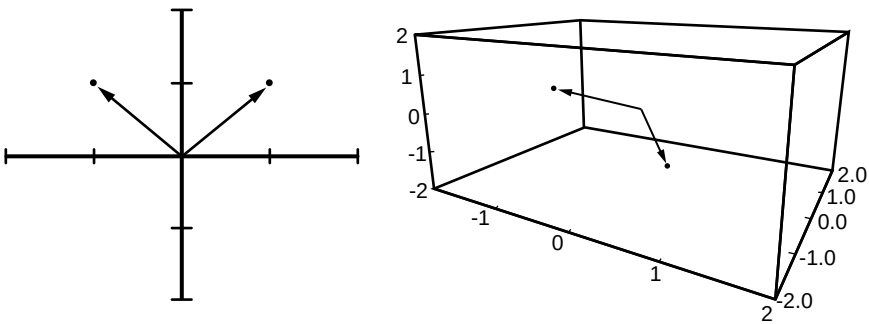

Figure 3.5 Vectors in two-dimensional and three-dimensional spaces that can be rotated to perfect agreement

As alluded to previously, however, Procrustes rotation is not generally associative when the dimensions of the matrices are not equal. Nonetheless, this example clearly shows that matrices of different dimensions can be rotated to maximal agreement provided they are matched with regard to the number of rows or the number of columns. In this case, the numbers of rows are matched.

BINARY PROCRUSTES ROTATION

Pairs of deep structure matrices, like other two-dimensional matrices, can also be rotated to agreement using the Procrustes transformation. In observation oriented modeling, the observed people, animals, behaviors, etc. will always be assigned to the rows of the deep structure matrices, and the units of ordered observations will always be assigned to the columns. The goal of the rotation, then, is to align the columns (units) in such a way that the co-occurrences of 1's in the two deep structure matrices are maximized. Although the previously discussed techniques, based on a singular-value decomposition of ${}_m\mathbf{X}'_p\ {}_p\mathbf{Y}_v$, could be used to accomplish this goal, the binary nature of all deep structure matrices permits a more efficient technique referred to as binary Procrustes rotation. This simplified transformation does not require computation of eigenvalues and eigenvectors, which could become unwieldy with large deep structures. It will also preserve the original scaling of the deep structure matrix to the extent the transformed values will range from 0 to 1. This feature will prove particularly useful in observation oriented modeling.

Binary Procrustes rotation begins by designating one deep structure as the target matrix and the second as the conforming matrix. The transpose of the conforming matrix, ${}_m\mathbf{X}'_p$, is then multiplied by the target matrix, ${}_p\mathbf{Y}_v$, resulting in the needed transformation matrix, ${}_m\mathbf{T}_v$,

$$ {}_m\mathbf{T}_v = {}_m\mathbf{X}'_p\ {}_p\mathbf{Y}_v. \qquad [3.4] $$

The transformation matrix is then normalized by columns and then normalized by rows. Normalization entails summing the squared values in the individual columns (or rows) of ${}_m\mathbf{T}_v$ and then dividing each value in the matrix by the square root of its corresponding column (or row) sum. The purpose of these normalizations is to ensure the values in the final rotated matrix will not exceed 1 and to reduce the impact of large discrepancies in the numbers of observations for the different units. Finally, the normalized

transformation matrix is multiplied to the conforming matrix to compute the rotated deep structure, ${}_{p}\mathbf{B}_{v}$,

$$ {}_{p}\mathbf{B}_{v} = {}_{p}\mathbf{X}_{m}\ {}_{m}\mathbf{T}_{v}. \qquad [3.5] $$

Consider, for example, 10 people who, like Ben and Erika, rate Cereal A on 7-point scales. The biological sex of these people is also noted, and the ordered observations and deep structures are recorded in Table 3.1. Designating biological sex as the target matrix, ${}_{10}\mathbf{Y}_{2}$, and the cereal ratings as the conforming matrix, ${}_{10}\mathbf{X}_{7}$,

$$ \text{Gender} = {}_{10}\mathbf{Y}_{2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} $$

$$ \text{Cereal A Rating} = {}_{10}\mathbf{X}_{7} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. $$

Table 3.1 Biological sex and cereal rating deep structures for 10 persons

			Deep structure		Deep structure						
Person	Biological Sex	Rating	M	F	−3	−2	−1	0	+1	+2	+3
P_1	M	−3	1	0	1	0	0	0	0	0	0
P_2	F	2	0	1	0	0	0	0	0	1	0
P_3	M	−2	1	0	0	1	0	0	0	0	0
P_4	M	−2	1	0	0	1	0	0	0	0	0
P_5	F	3	0	1	0	0	0	0	0	0	1
P_6	F	3	0	1	0	0	0	0	0	0	1
P_7	M	−3	1	0	1	0	0	0	0	0	0
P_8	M	−2	1	0	0	1	0	0	0	0	0
P_9	F	2	0	1	0	0	0	0	0	1	0
P_{10}	F	3	0	1	0	0	0	0	0	0	1

Following the steps outlined previously,

$$ {}_7\mathbf{T}_2 = {}_7\mathbf{X}'_{10}\,{}_{10}\mathbf{Y}_2 = \begin{bmatrix} 2 & 0 \\ 3 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 2 \\ 0 & 3 \end{bmatrix}. $$

Examination of the transformation matrix at this point reveals the simple nature of the rotation. The values in the matrix indicate the frequencies for which 1's were found for the same people in the joint columns of the two deep structures. The value for $\mathbf{T}_{[11]}$, for instance, equals 2, which indicates two matches between the first column of ${}_{10}\mathbf{Y}_2$ (males) and the first column of ${}_{10}\mathbf{X}_7$ (cereal rated as -3), whereas the value for $\mathbf{T}_{[21]}$ indicates three matches between the second column in ${}_{10}\mathbf{Y}_2$ and the first column in ${}_{10}\mathbf{X}_7$. The zeros in the transformation matrix indicate no matches across people between the joint columns. If the goal is to maximize the number of column-wise matches or hits between the two deep structure matrices, then the transformation matrix will essentially reveal how the columns in ${}_{10}\mathbf{X}_7$ can be aligned, combined, or split to

match the columns in ${}_{10}\mathbf{Y}_2$. Following the next steps in the process, ${}_7\mathbf{T}_2$ is normalized by columns and then by rows,

$$
{}_7\mathbf{T}_2 = \begin{bmatrix} 0.55 & 0 \\ 0.83 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0.55 \\ 0 & 0.83 \end{bmatrix}, \quad {}_7\,\mathbf{T}_2 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix},
$$

and finally ${}_{10}\mathbf{B}_2$ is computed, the rotated conforming matrix, which can be compared to the target matrix, ${}_{10}\mathbf{Y}_2$,

$$
{}_{10}\mathbf{B}_2 = {}_{10}\,\mathbf{X}_7\,{}_7\mathbf{T}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}; \qquad {}_{10}\,\mathbf{Y}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}.
$$

In this example, the rotated matrix matches the target matrix, ${}_{10}\mathbf{Y}_2$, perfectly. In other words, cereal ratings were transformed to perfect conformity with biological sex. How was this result accomplished? Returning to Table 3.1, it can be seen that all of the males rated Cereal A using the scale points designated by −3 and −2, whereas all of the females chose scale points designated by +2 and +3. Their seven-unit deep structured ratings could therefore be reduced without ambiguity to their two-unit biological sex deep structure.

As noted previously, the rotation of ${}_{10}\mathbf{X}_7$ to ${}_{10}\mathbf{Y}_2$ is not necessarily equal to the rotation of ${}_{10}\mathbf{Y}_2$ to ${}_{10}\mathbf{X}_7$. In fact, what is the outcome of rotating the two-unit biological sex deep structure to the seven-unit cereal ratings deep structure? Computing the transformation matrix,

$$ {}_2\mathbf{T}_7 = {}_2\mathbf{Y}'_{10}\,{}_{10}\mathbf{X}_7 = \begin{bmatrix} 2 & 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 3 \end{bmatrix}, $$

yields the same result as previously derived, only transposed. Next, normalizing by columns and then by rows,

$$ {}_2\mathbf{T}_7 = \mathrm{n}\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}, $$

$$ {}_2\mathbf{T}_7 = \mathrm{n}\begin{bmatrix} .707 & .707 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & .707 & .707 \end{bmatrix}, $$

results in a final transformation matrix with values quite different from those for rotating cereal deep structure to biological sex deep structure. Moreover, the final rotated matrix, ${}_{10}\mathbf{B}_7$, does not match the target matrix, ${}_{10}\mathbf{X}_7$,

$$ {}_{10}\mathbf{B}_7 = \begin{bmatrix} .707 & .707 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & .707 & .707 \\ .707 & .707 & 0 & 0 & 0 & 0 & 0 \\ .707 & .707 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & .707 & .707 \\ 0 & 0 & 0 & 0 & 0 & .707 & .707 \\ .707 & .707 & 0 & 0 & 0 & 0 & 0 \\ .707 & .707 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & .707 & .707 \\ 0 & 0 & 0 & 0 & 0 & .707 & .707 \end{bmatrix}; \quad {}_{10}\mathbf{X}_7 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. $$

Examining the values in ${}_{10}\mathbf{B}_7$, it is as if each value of 1 in the original conforming matrix, ${}_{10}\mathbf{Y}_2$, was split evenly ($.707^2 = .50$) between two units (columns) of ${}_{10}\mathbf{X}_7$. In fact, due to the column and row normalization of the transformation matrix, the sum of squared values in the rows of ${}_p\mathbf{B}_v$ will always equal unity, and to the degree the values in ${}_p\mathbf{B}_v$ depart from 0 and 1, they indicate a corresponding degree of ambiguity, or inaccuracy, in rotating one deep structure to another. Furthermore, recalling that the rows of deep structure matrices correspond to objects under study (often people), the values in ${}_p\mathbf{B}_v$ can be interpreted from their perspective. The previous values in ${}_{10}\mathbf{B}_7$, for instance, were all equal to .707, indicating that for every male and female in the group, the biological sex deep structure units could not be perfectly conformed to the cereal rating deep structure units.

This last point can be clarified by returning to the original data and the scenario of rotating the seven-unit cereal ratings deep structure, ${}_{10}\mathbf{X}_7$, to conformity with the two-unit biological sex deep structure, ${}_{10}\mathbf{Y}_2$. Recall that ${}_{10}\mathbf{X}_7$ could be rotated to perfect agreement with ${}_{10}\mathbf{Y}_2$. Now suppose one change to the original ratings is made—namely the first person (male) in the data set rates the cereal as +3 rather than -3. The deep structure is therefore

$$
{}_{10}\mathbf{X}_7 = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}.
$$

Comparing this matrix to the original matrix, it can be seen that only the first row is different. In the original matrix, "1" is located in the first column (corresponding to −3), and here it is located in the final column. The transformation matrix is computed as

$$ {}_7\mathbf{T}_2 = {}_7\mathbf{X}'_{10}\,{}_{10}\mathbf{Y}_2 = \begin{bmatrix} 1 & 0 \\ 3 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 2 \\ 1 & 3 \end{bmatrix}, $$

and normalized by columns and then by rows:

$$ {}_7\mathbf{T}_2 = \begin{bmatrix} 0.30 & 0 \\ 0.90 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0.55 \\ .30 & 0.83 \end{bmatrix}, \quad {}_7\mathbf{T}_2 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ .34 & .94 \end{bmatrix}. $$

Finally, ${}_{10}\mathbf{B}_2$, the rotated conforming matrix, is computed and compared to the target matrix,

$$ {}_{10}\mathbf{B}_2 = {}_{10}\mathbf{X}_7\,{}_7\mathbf{T}_2 = \begin{bmatrix} .34 & .94 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ .34 & .94 \\ .34 & .94 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ .34 & .94 \end{bmatrix}; \quad {}_{10}\mathbf{Y}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}. $$

With this one change, the conforming matrix can no longer be rotated to perfect agreement with the target matrix. Examining the values in ${}_{10}\mathbf{Y}_2$ from the perspective of the persons, the first row represents the male whose rating was changed. The large differences between his target (1, 0) and rotated (.34, .94) deep structure unit values indicate that the rotation was relatively inaccurate for his ordered observations. In other words, in the context of the other nine people, his cereal deep structure could not be aligned with, or conformed to, his biological sex deep structure. The rotation was much more successful for the remaining people. The target deep structure unit values for three females (0, 1) were nearly matched in the rotated deep structure (.34, .94), and the target deep structure unit values for the remaining six persons were perfectly matched. Changing the one male's rating therefore did not affect the rotated deep structure unit values for all of the persons but, rather, only those females, in this example, who also chose +3 on the cereal rating scale.

Binary Procrustes rotation is thus an efficient method for transforming one deep structure matrix to conformity with another. More generally, it can be seen as a method for conforming one pattern of observations to another. Because the technique preserves the range of scaling in the original deep structures, individual ordered observations of the objects (or persons) under study can be evaluated according to the success of the transformation in obtaining conformity. As with Procrustes rotations based on singular-value decompositions, binary Procrustes rotation can also be used to rotate matrices that are not completely matched in dimensionality. As long as the two matrices are matched on at least one dimension (rows using the conventions presented previously), the technique can be applied. Finally, unlike the general form of Procrustes transformation, binary Procrustes rotation does not require complex mathematics and can be accomplished with simple algebraic matrix operations.

This page intentionally left blank

CHAPTER 4

Modeling with Deep Structures

Contents

Introduction 45
A Simple Experiment 46
Modeling the Experiment 47
Graphs Both Old and New 53
A Probability Statistic 57
Concluding Models and Summary 59

INTRODUCTION

Observation oriented modeling is based on deep structure matrices and their various operations described in Chapters 2 and 3. A simple example is used in this chapter as a vehicle for demonstrating its key components and for introducing a novel graphing technique and probability statistic. It is also shown that, against the grain of contemporary practice, observation oriented modeling shifts the focus of analysis away from computed aggregates such as means and variances onto the observations themselves. In other words, the focus is shifted to the people, specific behaviors, animals, things, events, etc. under investigation. Summary statistics and contemporaneous methods of presenting data graphically are relegated to a completely ad hoc role, or avoided altogether, and the estimation of population parameters is eschewed. In fact, with observation oriented modeling, the notions of sample, population, and statistical inference are relegated to a secondary role or completely avoided. The psychologist instead worries less about fulfilling untenable assumptions (e.g., normal population distribution and equal population variances) and thinks more about the patterns of ordered observations relative to a competing perspective of chance.

The examples in this chapter are presented largely to explicate the techniques of observation oriented modeling, contrasting them in a general way to modern methods. The philosophical, theoretical, and practical aspects of observation oriented modeling are developed more fully in chapters that follow. These subsequent chapters also provide more thorough critiques of modern thought and practices with regard to the statistical and research methodology of psychologists and other social scientists.

Observation Oriented Modeling
ISBN 978-0-12-385194-9, Doi:10.1016/B978-0-12-385194-9.10004-0

© 2011 Elsevier Inc.
All rights reserved.

A SIMPLE EXPERIMENT

A cognitive psychologist, through her experiences as a student and from her understanding of contemporary theories of memory, hypothesizes that loud continuous noise will hinder an individual's ability to store events in long-term memory. Following the techniques of experimental design codified in modern textbooks on research methods, she designates three groups: an experimental noise group, a music comparison group, and a control group. She then determines the methods for her study in which each participant will be asked to examine 30 pictures of different objects (e.g., a pencil, a keyboard, and a dog) on a computer screen that are presented at a high rate of speed (i.e., 500 milliseconds per picture). The participant will then be asked to complete a distraction task that entails solving a series of mazes for 5 minutes. After solving the mazes, the participant will be asked to recall from memory as many of the 30 previously reviewed objects as possible. The testing conditions for the three groups will be controlled so that they are as similar as possible except for the background noise in the testing room. The participants in the experimental group will perform all of their tasks, including the mazes and recall, while listening to noise (radio static) played continuously at 100 decibels. Participants in the music condition will perform all of their tasks while listening to Beethoven's Ninth Symphony played at 100 decibels, and persons in the control group will work in silence. The psychologist recruits 18 participants from undergraduate courses in psychology at her university, and she randomly assigns the participants to one of the three conditions when they take part in the study. She specifically expects the students in the noise condition to recall fewer items than participants in the Beethoven or control conditions, and she expects the two latter groups to recall similar numbers of items. Her theory is not mathematically formal enough to predict the exact number of items recalled for any of the persons in the groups, nor does it yield predictions of aggregate (e.g., mean) outcomes for the three groups.

Despite potential flaws in her methods, in the parlance of contemporary research design, her study is a "randomized controlled experiment" because she systematically manipulated some aspect of the environment (noise, music, and silence) while controlling others (e.g., ensuring the memory tasks are equivalent), and she randomly assigned the participants to the three groups in an effort to prevent individual differences from confounding her results. Of course, the goal of designing a randomized experiment is to shore up the claim that one variable causes (according to J. S. Mill's definitions)

another, and in this instance the three conditions are designated as the "independent variable," the cause, and the number of items successfully recalled as the "dependent variable," the effect.

MODELING THE EXPERIMENT

The results for the psychologist's study are reported in Table 4.1. Following the guidelines found in modern textbooks on statistical analysis, the psychologist could analyze her data using analysis of variance. She could also analyze her data using a series of independent samples *t*-tests or some other type of statistic, such as the eta measure of association. Regardless, she would view her sample of 18 individuals as representative of a larger population, and she would employ null hypothesis significance testing in an attempt to make inferences about population parameters. For instance, with analysis of variance, she could contrast null and alternative hypotheses based on population means. She could also estimate the magnitude of the relationship between her independent and dependent variables if she had in fact studied

Table 4.1 Ordered observations for group membership and items recalled

Person	Group	No. of items recalled[a]
P_1	N	17
P_2	N	19
P_3	N	15
P_4	N	16
P_5	N	17
P_6	N	18
P_7	B	22
P_8	B	25
P_9	B	21
P_{10}	B	23
P_{11}	B	24
P_{12}	B	24
P_{13}	C	25
P_{14}	C	26
P_{15}	C	22
P_{16}	C	25
P_{17}	C	24
P_{18}	C	23

B, Beethoven; C, control; N, noise.
[a]0, no items recalled; 30, maximum.

the entire population.[1] Her analyses and conclusions would all be based on aggregates (i.e., means and variances) derived from her specific observations.

With observation oriented modeling, however, the psychologist will construe the study differently by shifting her focus from "variables" to the ordering of the observations in the study. With regard to the recall task, the psychologist and the participants both played a part in the ordering of these observations. The psychologist was responsible for the overall structure of the observations, as she designed the task and specified the methods of numbering the observations in whole units from 0 (no items recalled) to 30 (all items recalled). The specific features of the computer program, the selection of the mazes to be solved, and of course the noise and music to be played at a certain decibel level, were all determined by the psychologist in an attempt to obtain observations with a particular pattern.[2] The participants contributed to the ordering of the observations by following the instructions (probably to varying degrees) of the psychologist, paying attention to the tasks, memorizing, self-reflecting, and earnestly recalling from memory the different items. With regard to assigning people to different groups (noise, music, and control), the psychologist alone was responsible for the ordering of these observations. At the conclusion of each structured social interaction, the psychologist catalogs the ordered observations as shown in Table 4.1.

The psychologist is now, in a sense, staying as close to the reality of her observations as possible. She is also avoiding aggregates because her theory, like most all psychological theories, does not explicitly (i.e., via an integrated, causal model; see Chapter 6) call for their computation. Instead, the explanatory power of her particular theory rests on the individual memories of the persons in her study. The goal of her analysis is therefore to corroborate her hypotheses with regard to the persons in the study, not with

[1] The null hypothesis, for example, would be expressed in terms of population parameters as H_0: $\mu_{noise} = \mu_{music} = \mu_{control}$, where the Greek letter mu refers to the mean number of items recalled computed from the entire population of persons. In most research, the population parameters can only be understood theoretically because they have no empirical basis. The estimated magnitude of the relationship between the independent and dependent variables in the population would be most appropriately computed as ω^2.

[2] Expectancy effects have long been documented in which researchers inadvertently influence the outcomes of their studies in ways consistent with their biases (Rosenthal & Rubin, 1978). These specific confounding factors are not being referred to here; rather, the reference is to the explicit conscious decisions made by the researcher when designing her study.

regard to an aggregate. In other words, her theory is directly tied to the reality of each person's psyche, not to a nonexistent "average psyche" computed mathematically at the conclusion of her study. Observation oriented modeling allows her to honor and remain consonant with this fact.

The psychologist next acknowledges that her sample of 18 university students is not necessarily representative of the population. In fact, consistent with the nearly unanimous practice of psychologists, she has not even stated from which population she has sampled and to which she would generalize under null hypothesis significance testing—all undergraduates at her university, all Americans aged 18–20 years, all people in the world? Such notions are recognized as important for polling in an election or for purposes in which one is truly attempting to estimate a hypothetical population parameter such as a proportion, mean, or variance. With observation oriented modeling, however, the psychologist is not primarily interested in these aggregates, even if they are computed for her 18 students, much less some unspecified and hypothetical population. The psychologist is instead specifically interested in knowing if the ordered observations for the item recall task are essentially convertible with the ordered observations comprising group membership. In the language of observation oriented modeling, can the item recall ordered observations (the effect) be brought into conformity with the group membership ordered observations (the cause)?[3] Because ordered observations are represented in deep structure form, the question is whether the 31-unit item recall deep structure can be rotated to conformity with the 3-unit group membership deep structure. Recall from Chapter 3 that deep structures can be transformed to maximum agreement using binary Procrustes rotation, and the success of transforming one deep structure to another can be judged at the level of the observations in the analysis.

[3] It is noteworthy that Aquinas can be found speaking in general terms that effects conform to their causes. For instance, "Now it has been stated above that acts are called human, inasmuch as they proceed from a deliberate will. Now the object of the will is the good and the end. And hence it is clear that the principle of human acts, insofar as they are human, is the end. In like manner it is their terminus: for the human act terminates at that which the will intends as the end; *thus in natural agents the form of the thing generated is conformed to the form of the generator* [italics added]" (Aquinas, 2008, Ia 2æ, 1, 3). Also, "Further, no effect hinders its cause, since it is conformed to its cause" (Aquinas, 2008, Ia 2æ, 48, 1). This quote is taken from an objection, but St. Thomas does not disagree with this part of the objection in his reply. The notion of effects conforming to causes can also be seen in St. Thomas' view of the whole of ordered creation conforming to the intention of its principal cause, God.

Much more will be said about the differences between aggregate-based null hypothesis tests and the current observation oriented methods in later chapters, but it is important to note here that the psychologist's scientific desire to generalize beyond her 18 students is satisfied not through an appeal to population parameters but, rather, to replication and theory. A hallmark of science is repeatability, and the first action taken by any scientist who believes to have discovered one of nature's secrets is to conduct the exact same study a second time. The scientist will then anxiously await new replications performed in the laboratory and by scientists from other laboratories. Any experiment is framed by theory, and generalization is also accomplished via deduction and induction as the scientist both specifies and discovers the causes for the repeatable observations at the level of the observations (again, unless the model under consideration specifically and formally incorporates aggregates). Suppose, for instance, the psychologist is successful in rotating the item recall deep structure to perfect conformity with the group membership deep structure, and the pattern of results for each person—the level of observation—matches her hypotheses perfectly. Because she has studied only 18- to 20-year-olds, she must be careful not to generalize beyond this age group unless her theory specifically dictates otherwise. If her theory does not limit her capacity to generalize to, for example, 60- to 65-year-olds, then she will wish to empirically test this particular unconditional clause of her theory. She thus wins the right to generalize beyond her 18 students through replication, patient theoretical developments, and carefully designed empirical studies.

Having construed the immediate goal as one of rotating deep structures to maximum conformity, the psychologist is ready to analyze the ordered observations. The data in Table 4.1 are first converted to their deep structures. Group membership is treated as the target deep structure, the item recall scores are treated as the conforming deep structure, and binary Procrustes rotation is then applied (see Chapter 3). The rotated matrix, ${}_{18}\mathbf{B}_3$, and the group membership deep structure, ${}_{18}\mathbf{Y}_3$, are reported in Figure 4.1. The rows of each matrix correspond to the 18 participants in the study, and the columns (units) correspond to the three groups. The unit values in ${}_{18}\mathbf{B}_3$ for the first six people match their unit values in ${}_{18}\mathbf{Y}_3$ perfectly. The item recall deep structures for these individuals—considered in the context of the other 12 people—could be unambiguously conformed to their group membership deep structures. Two other individuals showed similarly clear results, whereas the rotated deep structure values for the remaining 10 persons were not as clear-cut.

$$
{}_{18}\mathbf{B}_3 = \begin{bmatrix}
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
0 & .71 & .71 \\
0 & .45 & .89 \\
0 & 1 & 0 \\
0 & .71 & .71 \\
0 & .89 & .45 \\
0 & .89 & .45 \\
0 & .45 & .89 \\
0 & 0 & 1 \\
0 & .71 & .71 \\
0 & .45 & .89 \\
0 & .89 & .45 \\
0 & .71 & .71
\end{bmatrix}; \quad
{}_{18}\mathbf{Y}_3 = \begin{bmatrix}
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 0 & 1
\end{bmatrix}
$$

Figure 4.1 Rotated conforming and target deep structure matrices for 18 participants

The results in ${}_{18}\mathbf{B}_3$ are specifically utilized in two ways in observation oriented modeling. First, recall from Chapter 3 that the values for each unit in ${}_{18}\mathbf{B}_3$ will range from 0 to 1, and when they conform with the target matrix, they will be strictly binary in accord with its deep structure (e.g., "1 0 0," "0 1 0," or "0 0 1"). The magnitude of the largest unit value in such cases will therefore be 1. To the extent that the row values in ${}_{18}\mathbf{B}_3$ cannot be rotated to perfect conformity to the target deep structure, the largest value in each row will be less than 1 and at least one other unit value will be greater than 0. Examination of the rows of ${}_{18}\mathbf{B}_3$ in Figure 4.1 reveals various examples. As already noted, the first six observations (persons) matched the target matrix perfectly (deep structures = "1 0 0"), whereas other observations were more ambiguous (e.g., 0, .45, and .89), and still others showed even splits between two of the three units (e.g., 0, .71, and .71). The largest value in each row of ${}_{18}\mathbf{B}_3$ can therefore be used as an indicator of the extent to which the rotation clearly discriminated between the units for each observation. This largest value is referred to as the classification strength index (CSI).

The second use of the rotated values in ${}_{18}\mathbf{B}_3$ is, in fact, classificatory. The target deep structure matrix is composed of strictly binary values, and in an

ideal outcome the values in the rotated conforming matrix will match these values perfectly. In other words, for each observation, the conforming ordered observations would be judged as convertible with the target ordered observations. Although the CSI reveals the extent to which the rotation discriminates between units for each observation, it does not indicate the similitude between the rows of the target and rotated matrices. Consider the eighth observation (row) in Figure 4.1, for instance, where the ranked rotated values of 0, .45, and .89 did not match the ranked values of 0, 1, and 0 in the target deep structure matrix. Although the CSI value for this observation is high (.89), it is located in the third unit, incorrectly suggesting the person participated in the control condition of the study.

The easiest method for judging such similitude between the rotated and target values for all of the observations is to dichotomize the values in ${}_{18}\mathbf{B}_3$ and compare to the values in ${}_{18}\mathbf{Y}_3$. As can be seen in Figure 4.2, this is done by replacing the highest value in each row of ${}_{18}\mathbf{B}_3$ with 1 and replacing the remaining row values with 0's. In the case of a tie for the maximum, all high

$$
{}_{18}\mathbf{B}_3 = \begin{bmatrix}
\mathbf{1} & 0 & 0 \\
\mathbf{1} & 0 & 0 \\
\mathbf{1} & 0 & 0 \\
\mathbf{1} & 0 & 0 \\
\mathbf{1} & 0 & 0 \\
\mathbf{1} & 0 & 0 \\
0 & 1 & 1 \\
0 & 0 & 1 \\
0 & \mathbf{1} & 0 \\
0 & 1 & 1 \\
0 & \mathbf{1} & 0 \\
0 & \mathbf{1} & 0 \\
0 & 0 & \mathbf{1} \\
0 & 0 & \mathbf{1} \\
0 & 1 & 1 \\
0 & 0 & \mathbf{1} \\
0 & 1 & 0 \\
0 & 1 & 1
\end{bmatrix}; \quad
{}_{18}\mathbf{Y}_3 = \begin{bmatrix}
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 0 & 1
\end{bmatrix}
$$

Figure 4.2 Dichotomized, rotated conforming matrix and target matrix for 18 participants. Unit values that match in both matrices are shown in bold in the dichotomized rotated matrix.

values are replaced with 1's and the remaining row values are again replaced with 0's. The 1's shown in bold in Figure 4.2 represent points of equality between the rows of the dichotomized matrix, ${}_{18}\mathbf{B}_3$, and the target deep structure, ${}_{18}\mathbf{Y}_3$. Overall, 12 of the 18 people were correctly classified, 6 of whom were in the noise group. The participants in the other two groups, however, were not all correctly classified, and 4 people could not be unambiguously classified to one group or the other. A summary of the classification results and CSI values can be presented in a number of ways. The specific output from the Observation Oriented Modeling software is shown in Figure 4.3. It can be seen that, overall, the psychologist was able to classify 66.67% ($12/18 \times 100$) of the persons correctly, which is reported as the percent correct classification (PCC) index. All 6 of the individuals in the white noise condition were classified correctly, whereas only 6 of the 12 individuals in the other two groups were classified correctly.

In summary, the number of items recalled (the effect) and group membership (the cause) ordered observations were converted to their deep structures. The effect deep structure was then rotated to maximum conformity with the cause deep structure. In other words, the 31-unit effect deep structure was reduced to a 2-unit deep structure on the basis of the patterns of ordered observations. These 2-unit conformed deep structure observations were then compared to the actual 2-unit cause deep structure observations, and the number of matches was tallied. The CSI values, which indicate the clarity in matching the conformed observations to the original observations, were also derived for the analysis. Overall, the results were somewhat favorable in light of the psychologist's hypothesis but were far from impressive.

GRAPHS BOTH OLD AND NEW

It is a cliché that happens to be true—"a picture is worth a thousand words"—and scientists, statisticians, politicians, and marketers have used it to their advantage for quite some time.[4] In observation oriented modeling, any established graphing technique is recommended for purely descriptive purposes. For instance, frequency histograms, pie charts, or stem-and-leaf displays can all be profitably employed as long as they do not lure the researcher into thinking inferentially about population parameters or

[4] Edward R. Tufte's (2001) classic book *The Visual Display of Quantitative Information* is highly recommended.

Model Tested :

```
  Condition  -->  Items Recalled
```

Classification Results

```
             Missing Observations  :  0
      Unclassifiable Observations  :  0
         Classification Precision  :  0

        Classifiable Observations  :  18
        Ambiguous Classifications  :  4
          Correct Classifications  :  12
   Percent Correct Classifications  :  66.67
```

Classification Summary by Condition

	Correct	Incorrect	Ambiguous
White Noise	6	0	0
Beethoven	3	1	2
Control	3	1	2

Note. Values represent totals. 18 observations classified.

Individual Classification Results

	Classification Result	Classification Strength	Target Deep Structure	Classified Deep Structure	Conforming Deep Structure
obs_1	C	1.00	White Noise	White Noise	17
obs_2	C	1.00	White Noise	White Noise	19
obs_3	C	1.00	White Noise	White Noise	15
obs_4	C	1.00	White Noise	White Noise	16
obs_5	C	1.00	White Noise	White Noise	17
obs_6	C	1.00	White Noise	White Noise	18
obs_7	A	0.71	Beethoven	Amb	22
obs_8	I	0.89	Beethoven	Control	25
obs_9	C	1.00	Beethoven	Beethoven	21
obs_10	A	0.71	Beethoven	Amb	23
obs_11	C	0.89	Beethoven	Beethoven	24
obs_12	C	0.89	Beethoven	Beethoven	24
obs_13	C	0.89	Control	Control	25
obs_14	C	1.00	Control	Control	26
obs_15	A	0.71	Control	Amb	22
obs_16	C	0.89	Control	Control	25
obs_17	I	0.89	Control	Beethoven	24
obs_18	A	0.71	Control	Amb	23

Note. C = Correctly Classified, I = Incorrect, A = Ambiguous.

Figure 4.3 Output from Observation Oriented Modeling software for 18 persons

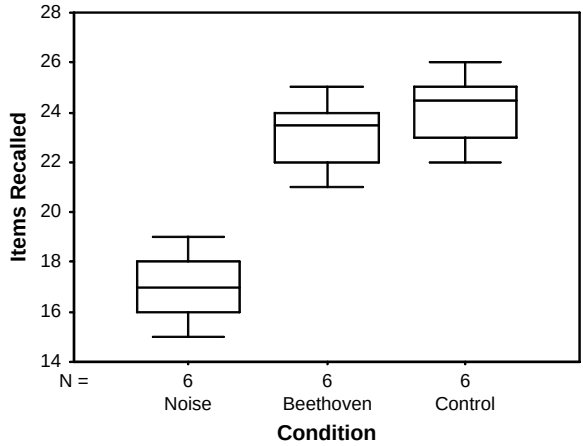

Figure 4.4 Box plot for all 18 participants in the three groups

forgetting about important assumptions (e.g., continuous quantitative measurement; see Chapter 8). A box plot for the psychologist's 18 participants is shown in Figure 4.4 and reveals the clear separation between the noise group distribution and the music and control group distributions, which are themselves highly similar.

A new method for graphing ordered observations that is similar to the box plot is presented here for observation oriented modeling based on deep structure matrices. This novel technique is referred to as a multi-unit frequency histogram, or simply multigram, and an example constructed for the current data can be seen in Figure 4.5. A multigram is created by using the units of the target deep structure to form columns and the units of the conforming deep structure to form rows. The widths of the columns in the graph are all set to an equal value that is determined from the largest frequency for a joint column and row in the entire data set. For the psychologist's ordered observations, for example, two people scored 17 in the noise condition, 24 in the Beethoven condition, and 25 in the control condition. Two is the maximum value observed, and the column width in Figure 4.5 thus indicates a frequency of two observations (persons). Separate frequency histograms for each column (i.e., unit of the target deep structure) are then generated and graphed, and the bars of the histograms are shaded (or color-coded in the software) to correspond to the classification results shown in Figure 4.3: Dark gray bars indicate successes, diagonally lined bars indicate failures, and white bars indicate ambiguous results. The overall and specific accuracy of the binary Procrustes rotation and classification results

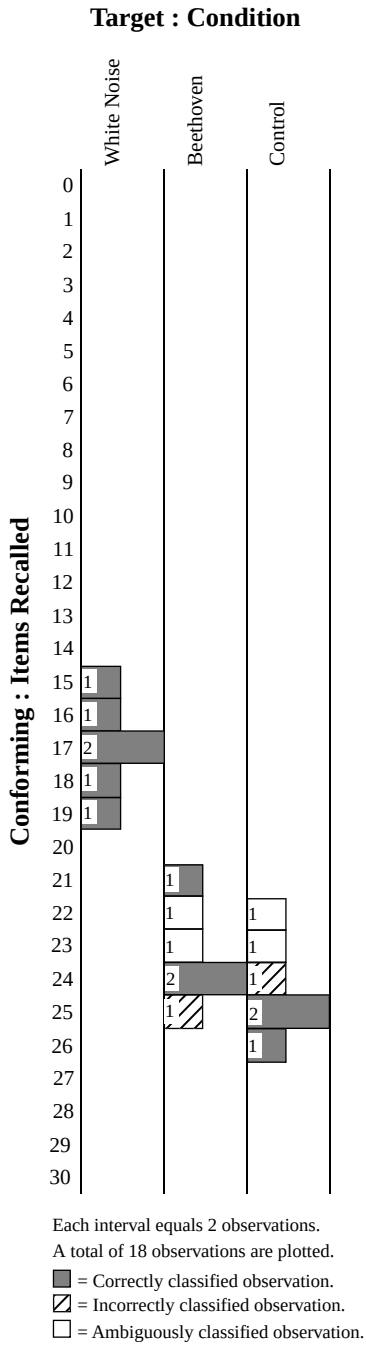

Figure 4.5 Multi-unit frequency histogram (multigram) for all 18 participants in three groups

can thus be evaluated quickly and efficiently in the multigram. The results in this example show clear support for the psychologist's overall prediction regarding the patterns of ordered observations, namely the participants in the noise condition recalled fewer items than the persons in the other two conditions. Moreover, the participants in the music and control conditions recalled similar numbers of items and therefore could not be differentiated by this particular analysis. The similarity between the music and control groups can be seen in their overlapping histograms with numerous ambiguous (white bars) and incorrect (diagonally lined bars) classifications in Figure 4.5.

A PROBABILITY STATISTIC

Considering the successful and unsuccessful classifications of the analysis raises an important issue, namely that of the psychologist projecting her biases into the interpretation of the pattern of results. In other words, from her perspective, the 66.67% accuracy reflected in the PCC index is likely to be interpreted as an impressive outcome, particularly because it appears graphically in the large number of white frequency bars in Figure 4.5. From the perspective of a critic, however, the 66.67% may seem highly probable under conditions of pure chance. The critic may argue that by repeatedly randomizing the 18 item recall scores, performing the binary Procrustes rotation, and classifying the outcomes, results of at least 66.67% accuracy would be fairly common. Consequently, the psychologist's results are judged as unimpressive; what looks to be an organized pattern of ordered observations (as explained by her theory) is more parsimoniously seen as a random outcome. Richard Harris has long argued that statistics are a form of social control over social scientists, and that significance tests prevent researchers from reading too much into their data:

> *For instance, it often proves difficult to convince a subject who has just been through a binary prediction experiment—involving, say, predicting which of two lights will be turned on in each of several trials—that the experimenter had used a random number table in selecting the sequence of events. Among researchers this tendency expresses itself as a tendency to generate complex post hoc explanations of their results that must be constantly revised since they are based in part on aspects of the data that are highly unstable from one application of the study to the next.*
>
> ***(Harris, 1994, p. 1)***

In this regard, probability statistics might serve as useful aids in resolving disagreements—such as the disagreement between the psychologist and her critic—regarding the interpretation of results.

With observation oriented modeling, the psychologist employs a probability statistic derived from what is commonly referred to as a randomization test. Recent developments in inferential statistics have successfully shown the general superiority of randomization tests and other resampling techniques, such as bootstrap and jackknife procedures, for robustly estimating population parameters.[5] Unlike the significance testing methods that have dominated the landscape for the past 70 or more years, these techniques rely less on mathematical theory and more on simulations of sampling processes performed on existing data sets or on data sets with defined distributional characteristics. Because observation oriented modeling is based on deep structure matrices, which all have a common form, the application of resampling methods is straightforward. Specifically, a randomization test can be conducted by first randomly shuffling the rows of the conforming deep structure. Binary Procrustes rotation is then performed on the original target deep structure and the randomized conforming deep structure, and the proportion of correct classifications is computed as described previously and recorded. This two-step process is repeated numerous (e.g., 500 or 1000) times, and the results are summarized and used as a base from which the actual proportion can be judged. For the cognitive psychologist's ordered observations, the results from such a randomization test indicated that 738 of 1000 trials (74%) produced proportions of successes at least as high as 66.67%. In other words, randomized versions of the psychologist's own data yielded PCC values at least this impressive ($\geq$66.67%) with unsettling frequency (.74).

The interpretation of this probability is notably post hoc and is used as an aid, as Harris suggests, to prevent the psychologist from interpreting her pattern of results too favorably because of a personal bias.[6] In addition, population parameters were not estimated nor was a null hypothesis

[5] For a summary of randomization tests, see Edgington and Onghena (2007). Robust statistical procedures, such as randomization tests, are also lauded by Erceg-Hurn and Mirosevich (2008).

[6] An early paper lauding the benefits of randomization tests is that of Winch and Campbell (1969). They wrote, "The method of randomization dispenses with the nagging problems of normality, probability sampling, etc. … We reason that it is very important to have a formal and nonsubjective way of deciding whether a given set of data shows haphazard or systematic variation. If it is haphazard, there is no reason to engage in further analysis and to be concerned about other threats to validity; if it is systematic, our outline shows that the analysis is not concluded with the test of significance but is just getting under way. And we believe it is important not to leave the determination of what is a systematic or haphazard arrangement of data to the intuition of the investigator" (p. 143).

entertained in this process, although it could be argued that the competing view represented by the randomization test is one of chance; that is, the same two sets of ordered observations, paired without any rhyme or reason, can be rotated to high agreement (at least 66.67%) with high frequency (74% of 1000 randomized sets).[7] In other words, chance shuffling of the observations yields just as much order as her theory in most cases, a conclusion that supports the critics' argument. Because of its origin in simulations based on randomized deep structures and its conceptual distinctiveness compared to the traditional *p* value employed in null hypothesis significance testing, this probability statistic is referred to as a chance value, or simply *c* value, in the realm of observation oriented modeling.

CONCLUDING MODELS AND SUMMARY

It is very important to realize that the PCC value, CSI values, and *c* value discussed previously were obtained from rotations of the full three-unit group (noise, music, and control) deep structure. The cognitive psychologist's hypotheses are more specific in nature, describing differences or similarities between only pairs of groups. She predicted that the participants in the noise condition would recall fewer items than participants in the Beethoven or control conditions, and she expected the two latter groups to recall similar numbers of items. Much like the distinction between the omnibus test and specific contrasts in analysis of variance, she must distinguish between the observation oriented model based on the three-unit group deep structure and models based on different two-unit deep structures that are more directly relevant to her hypotheses.

In order to specifically test her hypotheses, then, the psychologist conducts additional observation oriented analyses. First, she ignores the noise experimental group and performs the binary Procrustes rotation on the 2-unit (Beethoven and control groups) target deep structure and 31-unit item recall deep structure. The classification results in Figure 4.6 confirm what was noted previously; namely, the item recall deep structure could not be rotated clearly to the group deep structure. In other words, the model could not discriminate between the two groups with regard to the number

[7] Strictly speaking, the only assumption underlying the *c* value is complete independence. The units of observation are assumed to be independent both within and across different orderings when the randomization test is performed.

```
      Classification Results
              Missing Observations  :  0
       Unclassifiable Observations  :  0
          Classification Precision  :  0

         Classifiable Observations  :  12
         Ambiguous Classifications  :  4
           Correct Classifications  :  6
   Percent Correct Classifications  :  50.00

Randomization Results

  Observed Percent Correct Classified  :  50.00

           Number of Randomized Trials  : 1000.00
       Minimum Random Percent Correct  :   50.00
       Maximum Random Percent Correct  :  100.00
        Values >= Observed Proportion  : 1000.00
                         Model c-value  :    1.00
```

Figure 4.6 Classification and randomization test results for analysis comparing individuals in the Beethoven and control groups

of items they successfully recalled. The c value computed from 1000 randomized trials was 1.0, indicating that the proportion of at least .50 successful classifications could always be obtained from randomized orderings of the data. The multigram for this specific model would appear like the graph in Figure 4.5 but without the participants in the noise group.

The psychologist next conducts an analysis in which she combines the Beethoven and control conditions into a single group. The 31-unit item recall deep structure is again rotated to this new 2-unit (noise and combined groups) target deep structure, and the results are summarized (see Figure 4.7). Impressively, the model classifies all 18 participants correctly (PCC = 100%), and only 7 of 1000 randomized versions of the ordered observations achieve such a high proportion of accuracy ($c = 7/1000 = .007$). The multigram in Figure 4.8 shows the results in which all of the bars are dark gray, indicating perfect agreement between the rotated 31-unit item recall and 2-unit group deep structures. It can clearly be seen that every participant in the noise group recalled fewer items than every participant in the other two combined groups.

If she wishes, the psychologist can attempt to describe the "typical person" in each group using any aggregate statistic she deems most appropriate—for example, mean, median, or mode. Using the median, it

```
Classification Results

              Missing Observations  :  0
       Unclassifiable Observations  :  0
          Classification Precision  :  0

         Classifiable Observations  :  18
         Ambiguous Classifications  :  0
           Correct Classifications  :  18
   Percent Correct Classifications  : 100.00


Randomization Results

  Observed Percent Correct Classified  :  100.00

          Number of Randomized Trials  : 1000.00
       Minimum Random Percent Correct  :   72.22
       Maximum Random Percent Correct  :  100.00
        Values >= Observed Proportion  :    7.00
                        Model c-value  :    0.01
```

Figure 4.7 Classification and randomization test results for observation oriented model comparing individuals in the noise group with persons in the combined Beethoven and control groups

can be seen that the typical person in the noise condition recalled 17 items, whereas the typical person in the other two groups recalled 24 items. She can also speak of "effect size" in an aggregate manner, stating that the noise group median was seven items fewer than that of the other two combined groups. She is not obligated, however, to discuss her results in such aggregate terms; indeed, from the perspective of observation oriented modeling, she is encouraged not to do so. The psychologist can instead focus on the level of observation in her study. The description presented in Figure 4.9, using the American Psychological Association's technical style, demonstrates how she might report her findings.

In conclusion, what has the psychologist gained by using observation oriented modeling? The answer to this question is essentially a summary of what has been written previously. With observation oriented modeling, she places her focus where it belongs according to her theory and according to the operation of causes—namely on the persons in her study. Moreover, by

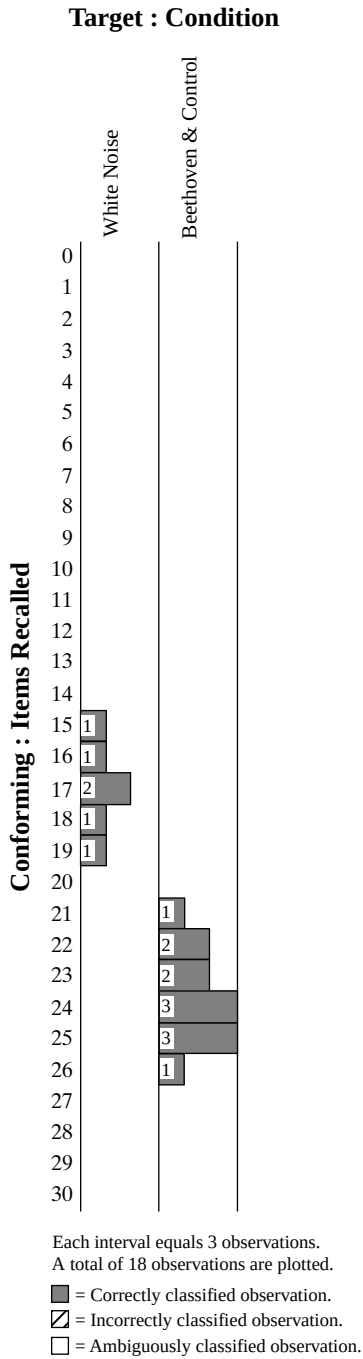

Figure 4.8 Multigram for model comparing individuals in the noise group with persons in the combined Beethoven and control groups

Observation oriented modeling was used to analyze the data. As expected, results indicated that the persons in the Beethoven and control groups were not clearly differentiated with regard to the number of items recalled (*c* = 1, 1000 randomized trials). Only 50% of these twelve participants were correctly classified by the analysis (median Classification Strength Index, CSI, = .89). Also consistent with our hypotheses, the persons in the combined Beethoven and control groups were clearly distinguished from the six persons in the noise condition *(c* = .007, 1000 randomized trials; all CSI values = 1). In fact, each person in the noise condition recalled fewer words than each of the other twelve persons in the Beethoven and control groups. The greatest number of words recalled in the noise condition was 19, while the fewest number of words recalled in the other two conditions was 21. Complete descriptive statistics are reported in Table 1.

Table 1. Descriptive statistics for number of items recalled (0 to 30) by the three groups

	Mdn	Min	Max	MAD
Noise	17.0	15	19	1.0
Beethoven	23.5	21	25	1.0
Control	24.5	22	26	1.0

Note. MAD = median absolute deviation from the group median.

Figure 4.9 Example report of results for observation oriented analyses using American Psychological Association style

working with deep structures, not only does she seek to make statements of efficient cause in her particular study but also she is acutely aware of its structural features that contributed to the patterns of ordered observations. In other words, instead of thinking primarily about "variables," she pays attention to the minute details of how the observations are brought into a particular order. As will be seen in later applications, this habit will translate into the practice of paying close attention to the "scaling" used in a given study and to the application of all four of Aristotle's causes. The psychologist also gains economy of thought with observation oriented modeling because she avoids the four scales of "measurement" as well as the notions of population and sample and the entire inferential statistical game. Gone are the distinctions between ratio, interval, ordinal, and nominal variables; issues of Type I, II, or III errors; statistical power; random sampling; and unrealistic distributional assumptions. A probability statistic is employed, but it is one that is computed from the psychologist's own ordered observations and serves a particular, well-defined role in the interpretation of results. Finally, she has a clear path to follow regarding her research, namely replication, delineation of framing conditions, and development of an integrated model (see Chapter 6) to explain the order of nature.

This page intentionally left blank

CHAPTER 5

Statistics and Null Hypothesis Significance Testing

Contents

Introduction 65
The Force of History 66
Null Hypothesis Significance Testing 74
Turning the Tide with Observation Oriented Modeling 83
Hybridized NHST 84

INTRODUCTION

Statistical methods, both descriptive and inferential, play a dominant role in a variety of modern sciences. It could be argued that in fact many sciences, such as psychology and sociology, stake claim to the label "science" based largely on their employment of statistical methods. The activity of researchers in these disciplines is marked by clever methods of assigning numbers to, for example, observations of overt behaviors or marked responses on self-report questionnaires. Vast numbers of such quantified observations are further gathered and submitted to the latest methods of statistical analysis endorsed by the various experts in the field. For instance, sophisticated techniques such as structural equation modeling, generalizability theory, or item response theory might be employed in an effort to model a set of observations. Underlying most of these techniques, simple and complex, is the notion of "statistical significance," which is held as the criterion for discriminating between results that are publishable and results that will be filed away, forever lost from the public eye. The ostensibly objective nature of these activities is moreover deemed as vastly superior to qualitative research, which is often chastised as pseudo-science. Indeed, as part of their training, students of psychology and the other social sciences spend countless hours studying an ever-growing number of data management routines and statistical analysis methods, usually in lieu of studying finite mathematics, calculus, and geometry with the majority of other natural scientists.

Observation Oriented Modeling
ISBN 978-0-12-385194-9, Doi:10.1016/B978-0-12-385194-9.10005-2
© 2011 Elsevier Inc.
All rights reserved.

Are these students spending their time wisely, and how did the virtual monopoly of statistical methods come about in the first place? The former question actually hinges on the latter, for which an answer is provided in this chapter in the form of broad intellectual and historical brush strokes.[1] The ideas presented in this chapter, which are primarily philosophical in nature, provide a propitious framework for clearly differentiating between modern statistical methods of analysis and observation oriented modeling in subsequent chapters.

THE FORCE OF HISTORY

The origin of experimental statistics can be found in the works of Sir Francis Galton, who famously developed the methods of correlation and regression in the late 1800s in his efforts to understand the nature of inheritance. The popularization of experimental statistics as well as their widespread acceptance as legitimate methods of scientific investigation—which continues to this day—can be traced to the late nineteenth- and early 20th-century works of Karl Pearson and Sir Ronald Fisher.

By the late 19th century Charles Darwin's theory of evolution had thoroughly captured the imagination of academics and the social elite of Great Britain. Its ostensible congruity with the socialist philosophy and eugenics programs that had become fashionable at that time no doubt contributed to its appeal. The young and brash Pearson understood what was at stake regarding this new view of humanity that ran headlong against the Christian view that had long been a dominant feature of the English identity. Fascinated by his own Quaker ancestry, it appears he truly struggled on a deep and personal level with the disorder and disunity sewn into his own psyche by social Darwinism.[2] There was much in the social teachings of Christianity, particularly those of the Roman Catholic Church, that he found appealing and worthy of continued assent, but he understood they were fundamentally anathema to the teachings of the new "religion" of scientism peddled by the likes of Thomas Huxley and fueled by the

[1] A more explicit account of how statistical methods rose to dominance in the disciplines of psychology and education can be found in Kurt Danziger's *Constructing the Subject* (1990).

[2] Pearson's struggles are expertly chronicled by Theodore Porter in his 2003 book *Karl Pearson: The Scientific Life in a Statistical Age.*

philosophy of Auguste Comte. In the end, Pearson converted to the new religion and penned *The Grammar of Science* in 1892 as his testimony to the limitless power of science—as he defined it—to settle the questions of humanity and to cure its social ills.[3]

The title of the book mimicked John Henry Newman's *A Grammar of Assent*, published in 1870, in which the Anglican convert to Catholicism (and eventual Cardinal-Deacon, Beatified in September 2010) wrote of the limits of the formal reasoning that makes possible logic, mathematics, and natural science (Newman, 1870/1947). Loosely parallel to the anthropology of Thomas Aquinas, Newman did not confuse the abstractive powers of the active intellect with the whole person, nor did he get lost in the subjectivism that entraps many philosophers today. Instead, Newman sought to present human knowledge as first and foremost experiential—that is, grounded in reality. Knowledge is "lived" in a world of facts and is not strictly a mechanically constructed series of concepts that conform to syllogisms or other rules of reasoning. A person's sensations, emotions, imagination, and intuition are necessarily involved in his or her reasoning processes, which considered as a whole constitute the "illative sense":

> *I have already said that the sole and final judgment on the validity of an inference in concrete matter is committed to the personal action of the ratiocinative faculty, the perfection or virtue of which I have called the Illative Sense, a use of the word "sense" parallel to our use of it in "good sense," "common sense," a "sense of beauty"; and I do not see any way to go farther.*
>
> ***(pp. 262–263)***

A philosopher, or any other person, who thus denies the possibility of certitude falls victim to a part–whole confusion, taking language, syllogisms, and other conceptual systems of the intellect to be the whole of reasoning. For instance, a man in love with his wife will with certitude say "My wife is a wonderful person" because it is made on the basis of the entirety of his existential encounter with his wife. Thoughts, emotions, sensory images, intuition, memories, future dreams—everything that constitutes his very being is involved in such a judgment. It is, in the truest sense, deeply personal. If he thinks to himself, however, "My wife is a wonderful person because she has a lovely face, engaging personality, and

[3] *The Grammar of Science* was first published in 1892 and revised in 1900. The final edition was printed in 1911.

beautiful smile" in an attempt to demonstrate the truth of his original statement, then he has moved primarily into the intellectual realm of concepts in which certitude will be shaken because he is considering only parts of the whole experience of his wife. He has, in essence, transformed his original declaration of what he knows into an abstract conditional statement, "If my wife is lovely, engaging, etc., then she is wonderful." The consequence of this subtle movement in the husband's thinking exemplifies why Newman argued that language and formal systems of knowledge, including those employed by natural scientists, presuppose the illative sense:

> *This is what I have to say on formal Inference, when taken to represent Ratiocination. Science in all its departments has too much simplicity and exactness, from the nature of the case, to be the measure of fact. In its very perfection lies its incompetency to settle particulars and details. As to Logic, its chain of conclusions hangs loose at both ends; both the point from which the proof should start, and the points at which it should arrive, are beyond its reach; it comes short both of first principles and of concrete issues. Even its most elaborate exhibitions fail to represent adequately the sum-total of considerations by which an individual mind is determined in its judgment of things; even its most careful combinations made to bear on a conclusion want that steadiness of aim which is necessary for hitting it. As I said when I began, thought is too keen and manifold, its sources are too remote and hidden, its path too personal, delicate, and circuitous, its subject matter too various and intricate, to admit of the trammels of any language, of whatever subtlety and of whatever compass.*
>
> ***(p. 216)***[4]

A theoretical physicist and philosopher and historian of science, Pierre Duhem, writing at the turn of the 20th century, similarly drew important distinctions between the positivist nature of physics and the reasoning capacities involved in cosmology and metaphysics. Physics is a rule-bound, logically consistent, and sensory-dependent form of reasoning that, as Gödel's theorem later demonstrated for all formal systems, can never prove its own internal consistency: "No scientific method carries in itself its full and entire justification; it cannot through its principles alone explain all these principles. We should therefore not be astonished that theoretic physics rests on postulates which can be authorized only by reasons foreign

[4] Newman's arguments have found a more modern expression in Michael Polanyi's "tacit knowledge." See his 1968 paper titled "Logic and Psychology." For a comparison of Newman and Polanyi, see Moleski (2000).

to physics" (Duhem, 1906/1954, p. 293).[5,6] Duhem argued that physicists can develop theories that over time converge analogously with cosmological theories, but the sum of his writings in philosophy of science can be seen as an effort to draw careful lines between physics and metaphysics so that their individual autonomy may be preserved.

Pearson would attempt to erase all such lines and countered in his *Grammar* that no realm of thought or domain of human activity was outside the purview of natural science, which he explicitly defined as the collecting of facts and inductive discernment of correlations:

> *The classification of facts and the formation of absolute judgments upon the basis of this classification – judgments independent of the idiosyncrasies of the individual mind – essentially sum up the* aim and method of modern science.
> ***(Pearson, 1892/1957, p. 6; emphasis in original)***

[5] The quote is taken from a paper titled "The Physics of a Believer," which is included in the 1954 edition as an appendix. The paper was originally published in 1905 in the *Annals de Philosophie chrétienne*. Lastly, Duhem is referenced here because of his efforts to protect the sovereignty of both science and metaphysics, which is a matter separate from endorsing his overall philosophy of science.

[6] Since Duhem's early work, Stanley Jaki has pointed out that Gödel's theorem similarly applies to the so-called "theory of everything" of modern physics. See his book *The Relevance of Physics* (1966). In a chapter titled 'A late awakening to Gödel in physics', Jaki (2006) also chronicles how some modern physicists, most notably Stephen Hawking, are coming to realize the limits of physical theories. It is worth quoting Jaki extensively here: "Herein lies the ultimate bearing of Gödel's theorem on physics. It does not mean at all the end of physics. It means only the death knell on endeavors that aim at a final theory according to which the physical world is what it is and cannot be anything else. Gödel's theorem does not mean that physicists cannot come up with a theory of everything or TOE in short. They can hit upon a theory which at the moment of its formulation would give an explanation of all known physical phenomena. But in terms of Gödel's theorem such a theory cannot be taken for something which is necessarily true. Apart from Gödel's theorem, such a theory cannot be a guarantee that in the future nothing essentially new would be discovered in the physical universe which would then demand another final theory and so on. Regress to infinity is no answer to a question that keeps generating itself with each answer. Gödel's theorem means, among other things, that physicists who aim at reading God's mind will not succeed, because they cannot read their own minds in the first place. A physicist, Paul Davies, who writes a book with the title *The Mind of God* should be the object of pity and not the recipient of a prestigious prize for progress in religion. Gödel's theorem remains a serious assurance to all physicists that their minds will forever be challenged by ever fresh problems. With a recourse to logic they would also know what to think of efforts to derive the very specific constants of physics from nonspecific considerations. Insofar as mathematics works with numbers, it will remain steeped in specifics all of which raise the question: Why such and not something else? It is that question which keeps the mind awake, or rather is raised by minds not prone to slumber" (pp. 10–11).

Science furthermore renders philosophy obsolete, and metaphysics, attached to theology, is also seen as a harmful form of poetry that lulls humankind's inquisitive mind into a state of servile slumber. Even poetry itself "will not satisfy our aesthetic judgment" (p. 17), nor will art; indeed, the scientists' account of the universe is "the only one which can permanently satisfy the aesthetic judgment, for it is the only one which can never be entirely contradicted by our observation and experience" (p. 35). Contrary to Newman and Duhem, then, Pearson elevated science to a position that encompassed the entirety of human experience; in short, he made the scientific method his metaphysic (which he would deny holding, of course). However, Pearson went even further in making his metaphysical view conform to the scientific method. In a remarkable chapter titled "The Facts of Science," he concluded that although science claims the whole of the universe as its field, a scientist can never know any thing in the universe. All that a scientist, or anyone else for that matter, can be certain of are his or her sense impressions:

> *And yet, if such a reader will think over what happens when he knocks his elbow against the table, I think he will agree that it is the sense impressions of hardness, and perhaps of pain, which are for him the realities, while the table, as a "source of these sense impressions," is the shadow.… The* real *table lies for him in the permanent association of a certain group of sense impressions, and that the shadow table is what might be left were this group abstracted.*
>
> ***(p. 69; emphasis in original)***

The man who therefore states "My wife is a wonderful person," must subordinate this statement to Pearson's metaphysic and realize that what he truly means is that the sense impressions his mind identifies as having a certain "sameness" (i.e., contingency and correlation) are labeled as "wife," "person," and "wonderful." He must admit agnosticism when it comes to what his wife *is*, although he is certain that like the table on which he may bump his arm, she remains a shadow of his sense impressions. As Duhem points out, by way of this reasoning, a scientist willingly shackles himself in Plato's cave, convinced that only through careful study of the shadows can his freedom be won (Duhem, 1957, p. 299). Recall in Plato's allegory that people are prisoners in a cave who mistake shadows on the wall for the truth of reality. The tragedy of the cave is magnified when a prisoner who breaks free and comes to the truth that a deeper reality (i.e., Ideas, from Plato's perspective) exists beyond the shadows is considered. Plato asserts that if such a man were to return and attempt to free the others, who have grown comfortable with their imprisonment, he would likely be killed.

Pearson knew of Newman's *Assent* and the realist philosophy it presented, and he rejected it. Instead, he elevated the scientific method to the level of a metaphysic, a veritable way of life, in which one heroically walks a stoic path of self-sacrifice that winds not up the mountain of truth but down into a dimly lit cave full of willing prisoners gaping eternally at dancing shadows.

The irony here runs even deeper when Pearson's advocacy of a science free of metaphysical shackles forged by the authors of philosophical systems and religious creeds is considered. Unfortunately, this irony was lost on the intellectuals, academics, and social elite of the day, who lauded Pearson's book as the new "good news," likely responding to its appeals to socialism and eugenics reforms that were presented literally as "claims of science":

> *No degenerate and feeble stock will ever be converted into healthy and sound stock by the accumulated effects of education, good laws, and sanitary surrounding. … If society is to shape its own future—if we are to replace the stern processes of natural law, which have raised us to our present high standard of civilization, by milder methods of eliminating the unfit—then we must be peculiarly cautious that in following our own strong social instincts we do not at the same time weaken society by rendering the propagation of bad stock more and more easy.*
>
> ***(Pearson, 1957, p. 27)***

Albert Einstein is reported to have formed a reading group in 1902 that started with the study of Pearson, although this fact in no way implies Einstein finally agreed with Pearson's positivism.[7] The popularity of the book and its overwhelming effect, even on Pearson's own future work, was to raise science to the level of an authoritarian metaphysical system to which all other human endeavors were to submit. Because science itself was equated with the accumulation of facts (sense impressions) that were inferentially categorized according to their correlations and contingencies, statistics and probability were *ipso facto* elevated to the highest level of esteem:

> *Their values [correlation ratio and coefficient of contingency] enable us to determine by qualitative or quantitative classifications the link between any two phenomena in the universe. They form the basis of the newer outlook on nature, which measures the association between phenomena, and reduces causation and mathematical function to a special and extreme case of contingency.*
>
> ***(Pearson, 1957, p. 177)***

Regarding the long-term impact of Pearson's efforts, one need look no further than the overwhelming reliance on statistical methods in the

[7] Here, I draw heavily from Porter's 2003 biography of Pearson. He describes Einstein's reading group on page 7.

medical, biological, and social sciences at the turn of the 21st century as well as the obsession surrounding meta-analyses, which are the quintessence of sorting through mountains of statistical relationships with the hope of bringing order out of chaos.

Pearson's *Grammar of Science* alone, however, would not have been sufficient to bring about the widespread use of experimental statistics witnessed today because it provided relatively few specific methods and procedures that scientists could use in their research. Indeed, it took the later efforts of R. A. Fisher to invent and collate the tools of research design and statistical analysis that could be applied widely in numerous domains of inquiry. Among his most influential books were *Statistical Methods for Research Workers* (1930; the first edition was published in 1925) and *The Design of Experiments* (1937; the first edition was published in 1935). These two particular tomes were handbooks of research and data analysis, filled with example data sets, computational formulas, and useful tables that made them indispensable texts for the laboratories of biological and agricultural researchers. By the time *The Design of Experiments* reached its seventh edition in 1962, statistics had become the fundamental tool for researchers in the social sciences as well. In *Constructing the Subject* (1990), Kurt Danziger shows very clearly through his historical research how statistical methods rose to prominence in large part due to the testing industry that emerged in the early 1900s. Fascinating alliances between psychologists and the military and between psychologists and educators led to the triumph of the Galtonian school of thought over the competing Wundtian tradition. As alluded to previously, a direct line can be traced from Galton to Pearson and to Fisher; hence, Danziger's "triumph of the aggregate" is nothing less than the triumph of statistical methods. In psychology, pockets of resistance, particularly the behaviorists and comparative psychologists, fought against the ubiquitous application of statistical methods, but they could not stem the tide. Their arguments regarding the potential for drawing erroneous conclusions from aggregates (notably, means) are still relevant today, but they have been almost completely forgotten by contemporary psychologists, who are kept busy learning an endless array of novel statistical techniques that are continually presented as "the next best thing."[8]

[8] Danziger reviews some of the criticisms regarding aggregates. See pages 153–155 in *Constructing the Subject*. For a more recent treatment of this issue regarding personality psychology and psychological testing, see Lamiell (1997, 2009).

Given the influential history of psychological testing, then, Fisher is not alone responsible for the modern dependence on statistical methods. Even in his early texts, he drew upon the work of other pioneers and presented their important contributions. Pearson's formula for the correlation coefficient and Student's approach for comparing estimates of independent population means (i.e., the *t*-test) are just two examples. Today, the list of names and the vast arsenal of statistical methods is daunting for any student beginning his or her trek into the world of social science research methods, which includes measurement theories underlying testing. Still, there are three contributions from Fisher that cannot be gainsaid for their impact on modern practice: (1) the randomized controlled experiment, (2) analysis of variance, and (3) null hypothesis significance testing. These contributions, wedded to the grandiose, positivistic philosophy of science popularized by Pearson, helped grow statistical methodology into a virtual monopoly, determining the ways scientists approach their targets of investigation. For psychologists in particular, it can be argued that statistical methods have dictated their approach to the study of persons for more than 70 years.

Monopolies are often looked upon negatively because of their stymying effect on competition and creativity. Such characterizations may well fit the current Pearsonian–Fisherian monopoly. As alluded to previously regarding the behaviorists and comparative psychologists, a minority of social scientists and scholars have criticized this model. They have rightly noted that the Pearsonian–Fisherian approach toward collecting and analyzing data is based on assumptions that are almost certainly violated in genuine research. They have also accused the approach of suffocating a viable strategy for understanding human behavior based on mathematical modeling, and they have challenged its positivistic basis. The force of these attacks, however, has repeatedly been deflected or simply ignored by the advocates of the Pearsonian–Fisherian model and by the researchers and consumers who rely on statistical methods. The premise of the book you are now reading is that these various criticisms have been ineffective because they have not provided a sufficient analysis of the philosophical and practical limitations of the Pearsonian–Fisherian approach, and they have not provided a suitable alternative set of methods for analyzing data. Observation oriented modeling has already been introduced as a viable alternative to the Pearsonian–Fisherian approach, and its capabilities are more fully demonstrated in subsequent chapters. Of immediate concern, however, are the contributions made by Fisher listed previously. These contributions must be thoroughly reviewed in order to (1) understand the philosophical premises they embody,

(2) understand their limitations, and (3) demonstrate how they are commonly misunderstood and abused by researchers. A different course can then be effectively charted that entails a sound philosophical framework to support observation oriented modeling. A proper place to begin is with Fisher's null hypothesis significance testing because of its ubiquitous nature as well as the important role it plays in determining whether or not the results from a given study are considered publishable. In the following chapter, an example hypothesis test is presented, thoroughly critiqued, and then re-analyzed from the perspective of observation oriented modeling.

NULL HYPOTHESIS SIGNIFICANCE TESTING

Null hypothesis significance testing (NHST) underlies the overwhelming majority of statistical tests and analytical tools currently employed by psychologists and other research scientists. Indeed, "statistical significance" as determined through NHST has served as a gateway for publishing studies, particularly in the social sciences, for at least 70 years. Danziger (1990) writes,

> *By the end of the 1930s certain practices in regard to the use of statistics in psychology had become well-established. "Statistical significance" had become a widely accepted token of factual status for the products of psychological research. If a particular finding reached a conventionally fixed level of statistical significance, it was automatically received into the corpus of scientific psychological knowledge; if not, it was unlikely to be published, unless accompanied by other, "significant findings."*
>
> ***(p. 152)***

The origins of this crucial technique can be found in Fisher's (1937) *The Design of Experiments*, in which he introduced the null hypothesis and spoke of its central role in experimental research:

> *In relation to any experiment we may speak of this hypothesis as the "null hypothesis," and it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.*
>
> ***(p. 19)***

The null hypothesis is essentially a statement regarding a population parameter or difference between population parameters that provides the basis for determining the probability of an observed outcome. A population parameter is a value such as a mean, proportion, or index of variability that can be determined or computed, at least in principle, for a complete set of

observations. Fisher famously introduced NHST in the context of a lady attempting to determine, by tasting eight cups of tea, if the milk had been added first to the tea or if the milk had been added after the tea had been poured. Because the lady had declared she could discriminate the order, this was the "effect" attempting to be demonstrated. The null hypothesis was stated in contrast to this effect and posited that the lady was simply guessing and could therefore do no better than chance. The "experiment" was set up so that she would sort the cups into two sets of four (tea first or milk first). Assuming the lady was simply guessing, the probabilities associated with dividing the eight cups into two sets of four were then computed using rules of permutation. For instance, the probability of correctly sorting all eight cups correctly into two groups was .01 (1/70), provided four of the cups had the milk added first and the remaining four cups had the milk added second. The probabilities for three, two, one, and zero correctly sorted pairs were .23 (16/70), .51 (36/70), .23 (16/70), and .01 (1/70), respectively. With these probabilities in hand, the experimenter could determine a "cut-point" for declaring the result as "statistically significant." By the second edition of *The Design of Experiments*, Fisher suggested that an *a priori* or critical probability value (i.e., a critical *p* value) of .05 could serve as conventional cut-point for researchers. For the tea-tasting lady, this meant she would have to obtain a result that was likely to occur less than 5 times out of 100 assuming the null hypothesis was true; in other words, she would have to sort all eight cups correctly into the two groups (observed $p < .01$, 1/70). The probability of sorting three pairs of cups correctly, assuming she was guessing, was too high (.23, $p > .05$) and not extreme enough to prompt the experimenter to subjectively doubt the null hypothesis. With all of this groundwork laid, the experiment could finally be conducted by asking the lady to actually taste the eight cups of tea, sort them into the two groups, and then record and evaluate the result as statistically significant or as nonsignificant.

As a second example of NHST in *The Design of Experiments*, Fisher (1937) re-analyzed data collected by Darwin that were originally analyzed by Galton. The goal of the study was to compare the growth rates of cross-fertilized and self-fertilized plants that had been grown in environments that were made as equivalent as possible; for example, pairs of each type of plant were grown in the same pots. Fisher initially approached the null hypothesis from two slightly different angles. First, he defined it as a single, average limit that would be approached by both types of plants if the experiment were conducted indefinitely on additional pairs of plants. In other words,

the average heights of the cross-fertilized and self-fertilized plants would, in the long term, approach equality. Next, understanding the measure of height as a continuous quantity, Fisher drew upon the theory of errors and stated,

> *It is this circumstance that makes it appropriate to choose, as the null hypothesis to be tested, one for which an exact statistical criterion is available, namely that the two groups of measurements are samples drawn from the same normal population.*
>
> ***(p. 40)***

In other words, the null hypothesis posits one common population with a single mean height, and the cross-fertilized and self-fertilized plants are assumed to be randomly drawn from this population. One would expect, therefore, that mean heights computed from samples of both types of plants would tend to be equal. In fact, tying back into Fisher's first statement about the null hypothesis, one would expect to observe the means for the cross-fertilized and self-fertilized plants converging to a common value as samples are randomly drawn and combined. Incorporating the work of Student, Fisher finally described how the data were analyzed using the *t*-test, which essentially entailed comparing the actual mean heights computed from Darwin's data to a hypothetical distribution of means derived on the assumption of the null hypothesis being true. Like the tea-tasting "experiment," then, the probability of drawing Darwin's means randomly from this constructed distribution could be computed. If the observed means were improbably different, according to some standard cut-point such as .05, then the null hypothesis could be subjectively overturned. Indeed, Darwin's data proved to be "statistically significant," leading to the conclusion that the population mean heights for the cross-fertilized and self-fertilized plants were not equal.

What can be learned, in general, from these two examples of null hypothesis significance testing? The first thing that should strike any reader of Fisher's tea-tasting lady is its triviality. Perhaps this statement is too harsh because Fisher was using the example to demonstrate a set of techniques. On the other hand, the example sets the stage for the infamous Rhine studies on extrasensory perception, which have left the impression that NHST is more suitable for parlor games than for scientific exploration. More practically, it has left a lasting impression that replication is not an important aspect of scientific investigation. Imagine the tea-tasting study actually being conducted and the lady in fact sorting all eight cups correctly. Would any scientist be satisfied with the result? We can presume that Fisher would

surely answer "no" and demand at least a half-dozen replications. Suppose the woman succeeded in replicating her feat a dozen more times. The statistical probability of such an outcome, assuming she were in fact guessing, would be extremely low; but would anyone actually care to compute a probability at that point? If a researcher answers in the affirmative, he or she would necessarily agree that the tea-tasting game should go on indefinitely because it is only when the entire population of observations has been collected that a certain determination can be made regarding whether or not the lady was guessing. No researcher, of course, would adopt such an extreme stance or even demand a large number of replications (e.g., 25) for the tea-tasting lady; rather, at some point determined by his or her "illative sense," the researcher would move on to a different problem or conduct further studies to actually explain (via an integrated model; see Chapter 6) how the lady can accurately differentiate between the cups of tea. The point here is that NHST is not inherently "scientific" in the sense that it necessarily contributes to knowledge of nature through its causes.

Fisher's second example corroborates this disposition toward NHST, particularly when he quotes Darwin as stating that his experiment required 11 years to complete and would not likely be replicated in the near future. The entire goal of the re-analysis was thus to make some statement about the population mean heights of cross- and self-fertilized plants, implying the outcome would somehow obviate replication. Fisher never states this explicitly in reference to Darwin's data, but in *The Design of Experiments* he speaks of replication only in the context of increasing the number of levels in a factorial experiment, not in terms of running experiments repeatedly from beginning to end. As discussed later, the lack of attention paid to exact replications is partly due to the fact that NHST is a mathematical–probabilistic system that reflects Fisher's positivistic biases regarding the aims and nature of scientific investigation. The practical outcome for modern social scientists is the widespread mistaken notion that NHST is an adequate substitute for replication.

Another important lesson to learn from Fisher's two examples is perhaps more abstract than the issue of replication, but it is certainly no less important. Let's begin by more closely examining the re-analysis of Darwin's data and again ask why NHST was conducted. Fisher provided the answer succinctly in his *Statistical Methods for Research Workers* (1930): "The calculation of means and probable errors shows a deliberate attempt to learn something about that population" (p. 3). In other words, the goal was simply to estimate the difference between two population means. Fisher was

re-analyzing Darwin's data in 1935 and understood from the laws of inheritance that the average heights of the two types of plants should not drift toward some common, yet unspecified, population value; therefore, he stated the null hypothesis as one of equal population means. He computed the mean difference between pairs of plants as 20.93 inches and observed that the result was clearly different from zero but based on only 15 pairs of plants. It is at this juncture that Fisher, who was foremost a first-rate mathematician enmeshed in positivism, concerned himself primarily with the process of judging data through an abstract system of population parameters, probabilities, distributional assumptions, and random samples:

> *For a given population we may calculate the probability with which any given sample will occur, and if we can solve the purely mathematical problem presented, we can calculate the probability of occurrence of any given statistic calculated from such a sample.*
>
> ***(p. 9)***

In other words, NHST is at its core a mathematical–probabilistic system of thought. Using NHST for Darwin's quantitative data, Fisher (1) assumed the null hypothesis was true, (2) assumed a normal distribution for differences in heights for the entire hypothetical distribution of plants, (3) assumed the heights of the pairs of plants to be independent, (4) assumed a normal distribution of mean differences computed for samples drawn from the population, and (5) assumed the samples were drawn randomly from the population. Fisher did not take the time to spell out these various assumptions, but they are now well-known by students of statistical methods who have been taught the various assumptions of the dependent samples *t*-test Fisher used on Darwin's data. In any event, the continuous quantitative nature of height and the various assumptions provided the basis for which Fisher could finally compute the probability associated with randomly sampling 15 pairs of plants with a mean difference of at least 20.93 inches in height. The probability turned out to be slightly less than .05, and Fisher judged it as "statistically significant." The practical consequence of employing this abstract, and arguably convoluted, framework was to shore up Darwin's inference that the populations of cross- and self-fertilized plants differed with regard to their mean heights.

Compared to the abstract mental gymnastics involved in NHST, not to mention its numerous assumptions and probabilistic conclusions, the amount of knowledge gained from Darwin's data was meager. Paul Meehl understood as much when he starkly proclaimed that NHST is "one of the worst things that ever happened in the history of psychology" (1978,

p. 817). He went on to argue that those who use NHST are never required to engage in the formal and substantive theorizing that would lead to specific predictions, a point plainly seen in both of Fisher's examples. Darwin was not required to predict the exact difference in mean height between the cross- and self-fertilized plants, and the tea-tasting lady was not required to predict her proportion of successes before tasting the tea. A quick perusal of the literature from any particular domain of study will convince the reader that psychologists and social scientists rarely make specific point predictions as well. The basis for Meehl's arguments is the fact that NHST is at its core a mathematical–probabilistic scheme crafted within the philosophical context of positivism in which laws are abstracted from the correlations and contingencies of sense impressions. Positivism has long been discredited as providing an adequate description of scientific activity, and some of its opponents, most notably Karl Popper, have charged that such purely inductive philosophies can actually thwart progress when researchers hold their "frameworks" (or theories) in higher esteem than the real problems they are studying (Popper, 1996). In their day-to-day research, the activities of scientists can best be described as exhibiting a commonsense realist philosophy in which the goal is to understand the natures (i.e., their forms, powers, properties, etc.) of things at different levels—for instance, understanding molecules of water acting together to cause erosion compared to understanding water as a molecule composed of hydrogen and oxygen. Because NHST is best situated in a positivistic philosophy, it works against this proper attitude or disposition that constitutes the essence of scientific activity. As Meehl concludes, the result can only be a slow or nonexistent progress in scientific knowledge for those disciplines, like psychology, that rely heavily on NHST: "Theories in 'soft' areas of psychology lack the cumulative character of scientific knowledge. They tend neither to be refuted nor corroborated, but instead merely fade away as people lose interest" (Meehl, 1978, p. 806).

It is instructive to consider how Fisher's charge of data falsification against Gregor Mendel serves as an early contrast between the positivism inherent in NHST and the realist view of scientific activity. Recall that Gregor Mendel was an Augustinian priest living in the 19th century who famously studied pea plants in his monastic garden in Moravia. Conducting carefully controlled experiments, he self- and cross-fertilized the plants and meticulously recorded specific traits, such as length of stem, pod color, and flower color, across generations. Mendel conducted his experiments over

the span of 8 years and first published his results in an 1866 paper titled "Experiments on Plant Hybrids." Based on the consistent patterns visible in his data, Mendel derived the notions of dominant (which he referred to as dominating) and recessive traits as well as the segregation and independent assortment laws of inheritance. Somewhat tragically, Mendel's discoveries were largely ignored until the turn of the 20th century some years after his death in 1884. Three different scientists simultaneously rediscovered his important work, and today Mendel is recognized as the father of genetics. More tragically, Fisher accused Mendel of cooking (i.e., falsifying) his data in a famous paper published in 1936. Perhaps unsurprisingly, Fisher's accusations were largely based on a re-analysis of some of Mendel's results that, applying the principles of NHST, were deemed to be too improbable or simply "too good to be true." In other words, the results were judged as "statistically significant" in a negative manner that indicated a clear confirmation bias in Mendel's studies. More recent examinations of Mendel's experiments and data, however, have clearly shown that he did not cook his data. In fact, several scientists have shown that Fisher suffered from a strong confirmation bias that led to errors in his re-analyses. Unfortunately for Mendel, Fisher's accusations have stuck "like dirt sticks to a candidate after a mud-slinging political campaign," and he is still considered by some uninformed scholars to be a "betrayer of truth" and "rogue" (Hartl & Fairbanks, 2007, p. 976; see also Hartl & Orel, 1992; Monaghan & Corcos, 1985; Weiling, 1986).

Fisher developed a reputation for being intemperate in his disagreements with other scholars. In his review of Mendel's paper, he explicitly challenged William Bateson's attempt to use Mendel's findings in a critique against aspects of Darwinian evolution. Fisher may thus have been attempting to counter Bateson's arguments by dynamiting their cornerstone, namely Mendel's experiments. Regardless, it is clear that for Fisher, science is a mathematical–probabilistic enterprise like that seen in NHST. It is this framework that has priority over all else, and he used it like a sword to cut down his opponents:

> *Unfortunately he [Bateson] was unable to grasp the mathematical and statistical aspects of biology, and from this and other causes, he was not only incapable of framing an evolutionary theory himself, but entirely failed to see how Mendelism supplied the missing parts of the structure first erected by Darwin. Nothing but harm can come from following Bateson in regard to evolution theory.*
>
> ***(Fisher, 1929)***

Fisher also projected his view onto Mendel by arguing that the experimenting priest was primarily concerned with demonstrating a method instead of actually discovering something about nature:

> *If we read his [Mendel's] introduction literally we do not find him expressing the purpose of solving a great problem or reporting a resounding discovery. He represents his work rather as a contribution to the* methodology *of research into plant inheritance.*
>
> ***(as quoted in Mendel, 1964, p. 3; emphasis in original)***[9]

Mendel's introductory statements in his famous paper are clearly modest but suggest otherwise:

> *The striking regularity with which the same hybrid forms always reappeared whenever fertilization took place between the same species induced further experiments to be undertaken, the object of which was to follow up the developments of the hybrids in their progeny.*
>
> ***(p. 7)***

He goes on to discuss how no laws of inheritance have been satisfactorily stated to describe such regularities and that sufficiently detailed observations are not currently available. He then states that carefully controlled, lengthy studies such as the ones he is about to report are the only way to achieve the goal of finally reaching "the solution of a question the importance of which cannot be overestimated in connection with the history of the evolution of organic forms" (p. 8). Fisher's interpretation also does not square with Mendel's own language throughout his paper, in which he is perfectly comfortable speaking of traits, such as dominant and recessive traits, as properties of the plants themselves that explain the patterns in his observations. The laws of inheritance are therefore valuable because they help us to understand the nature of the pea plants, not because they are themselves mathematical and probabilistic formulations. Perhaps it speaks volumes that Mendel was able to discover these laws without the aid of NHST or other sophisticated statistical methods. Indeed, he relied primarily on permutations and ratios and at one point employed a simple mean to combine results from a replicated trial.

Fisher's misrepresentation of Mendel's overarching philosophy as well as his empirical results may be a clue that Meehl was correct when he stated, "Sir Ronald has befuddled us, mesmerized us, and led us down the primrose path" (1978, p. 817). Two additional facts attest to the veracity of Meehl's

[9] This edited volume presents Mendel's original paper, Fisher's original paper questioning Mendel's experiments, and Bateson's biographical note on Mendel.

colorful yet ominous statement. First, surveys have repeatedly shown that social scientists and educators who employ and teach NHST simply do not understand what they are doing. One of the earliest surveys found that a majority of psychologists could not accurately define what it meant to obtain a statistically significant result at the .05 or .01 level; in other words, they could not define the very criterion for which they determined whether or not a result was worthy of adding to the body of knowledge they considered scientific. One of the most common errors these psychologists made was to interpret a p value of .01 as indicating the study would successfully replicate 99% of the time if repeatedly conducted (as reported in Cohen, 1994). In other words, they viewed statistical significance as a suitable substitute for replication—a finding that confirms the argument made previously regarding the lessons learned from Fisher's original examples. A survey of German academics who teach statistics indicated that such errors still persist and are in fact handed down from generation to generation and perpetuated in leading textbooks (as reported in Gigerenzer, 2004). Pair these fundamental errors with the fact that statistical significance serves as the gatekeeper to publication in the social sciences, and a stunning and disheartening contradiction is revealed. Imagine a local town hiring a group of engineers to design and construct a new bridge. Further imagine the reactions of the council members upon learning that the engineers could not correctly define the mathematical and geometric concepts necessary to complete the project. Would the council retain the engineers? Those for whom NHST serves as the backbone of their research programs must honestly ask themselves the difficult analogous question regarding their own practices.

Second, social sciences tend to splinter and fragment over time rather than unify. Psychology is perhaps the prime example of this phenomenon. The Galtonian psychologists outmaneuvered and outlasted the Wundtian school, Adler and Jung broke away from Freud to develop their own theories, and a group of disgruntled "experimentalists" broke away from the "clinicians" of the American Psychological Association in the 1980s to form the American Psychological Society. In a famous address to his colleagues at the annual convention of the American Psychological Association in 1974, Lee Cronbach attempted to mend at least one fissure by highlighting the complementary nature of "experimental" and "correlational" research (Cronbach, 1975). His efforts could only be in vain, however, because the idealism underpinning positivism had already spread its roots deep into the conscious and unconscious minds of psychologists. Recall Pearson's

description of things as "shadowy unknowables behind that real world of sense impression in which we live" (1957, p. 73). The things of the mind derived from correlations and contingencies of sense impressions are primary, whereas the reality of the things themselves is illusory. Such a subjective attitude creates a propitious environment for breeding division because disagreements are handled by creating coteries rather than confronting problems anchored in an existing reality. The Galtonians viewed the Wundtians as inferior because of their methods, Adler and Jung broke from Freud because their models of the psyche were irreconcilable, and the experimentalists accused the clinicians of not being "pure" scientists. Today, the divisions in psychology are marked by superficial lines based entirely on subjective criteria meant to carve out and protect paradigms. Of course, the tragedy is that very few psychologists communicate across party lines because each group construes itself as superior to (or at least unrelated to) the next, even though they all study the same reality—the human person. Indeed, had psychologists adopted Aquinas' approach toward dividing sciences, they would have started from the reality of the person (i.e., from the essence of what a person *is*) rather than the conceptual models within their own psyches (Aquinas, 1963). Such is the fate, however, of those who adopt a Cartesian idealism, whether in the form of positivism or Kuhnian historicism with its incommensurable paradigms.

TURNING THE TIDE WITH OBSERVATION ORIENTED MODELING

What is needed, then, to overcome the force of history behind the current Pearsonian–Fisherian monopoly is a philosophical position and set of corresponding techniques that match the natural attitude and day-to-day activities of social scientists as they seek to understand persons. The modern student assumes that arming himself or herself with an arsenal of statistical techniques, including NHST, will yield the keys to science. The student will spend countless hours and inordinate mental resources learning how to accommodate assumptions—for instance, choosing just the right statistical test for the observations, transforming data to impose a normal distribution, throwing out observations because they are too discrepant from the mean, changing a theoretically informed model for the sole purpose of freeing up degrees of freedom, or struggling to find ways of correcting the non-Grammian properties of a correlation matrix. Upon completion of training, the student can later be found searching through data sets for those few

statistically significant (i.e., publishable) results, agonizing over whether or not an observed *p* value of .057 is truly nonsignificant, and spending months aggregating descriptive statistics from hundreds of articles into a single estimated "effect size." All of this activity is quite sophisticated, but as we have seen, it may be nothing more than a mathematical–probabilistic exercise that fails to produce viable scientific knowledge—that is, demonstrable knowledge of nature through its causes.

Observation oriented modeling is a technique that greatly simplifies how students and seasoned researchers approach their ordered observations (i.e., their data) from the philosophical position that recognizes the primacy of the real. These new techniques are relatively free of assumptions and therefore permit the scientist, like Mendel, to focus on patterns within the observations. The procedures used in a given study as well as the methods for obtaining the observations, such as questionnaires, measures of reaction time, or behavioral checklists, can consequently be more carefully considered for their role in the formal order of the observations. By rotating deep structures to maximal agreement, patterns in the ordered observations obtained from, for example, numerous persons or from a single person can be discerned and interpreted in light of particular hypotheses or an integrated model (see Chapter 6). Because the analyses involve deep structures, both quantitative and qualitative hypotheses can be evaluated, including hypotheses that entail logical statements. The researcher therefore need not make unwarranted assumptions about nature or his or her methods when using observation oriented modeling—for instance, assuming that psychological qualities such as intelligence or extraversion are structured as continuous quantities or assuming that participants in a study have been drawn randomly from a larger population. Instead, the researcher can grant primacy to reality and tailor his or her methods and analyses to best address the natures of the persons, animals, or things under investigation via material, formal, efficient, and final causes. The particular details of this realist philosophy are elucidated in subsequent chapters as example data sets are modeled using observation oriented modeling and compared to traditional statistical techniques.

HYBRIDIZED NHST

In closing this chapter, it is important to note that Fisher's original null hypothesis test has been greatly modified throughout the years. Modern students are in fact taught what is best described as a hybrid of ideas

developed through the intellectual, and often personally insulting, debates between Fisher and a pair of fellow statisticians, Egon Pearson (Karl Pearson's son) and Jerzy Neyman. In the 1930s and 1940s, the younger Pearson and Neyman challenged Fisher's approach toward significance testing, particularly noting its failure to include the researcher's specific hypothesis—that is, his or her expectations of positive results from the experiment. This "alternative hypothesis" for the tea-tasting lady, for instance, would have been the frequency that she expected to accurately sort the teas into two groups. Fisher never reconciled his disagreement with Pearson and Neyman, but their different views of significance testing have nonetheless been fused by modern authors of statistics textbooks. The particular contributions of Pearson and Neyman to this NHST hybrid can be clearly seen in considerations of Type I, Type II, and Type III conclusion errors as well as in considerations of statistical power. The details of this hybrid are discussed in greater detail in the next chapter utilizing a worked example. At this juncture, it is sufficient to note that the hybrid retains Fisher's original view of hypothesis testing as a method that revolves around estimating population parameters. Whether or not the hybrid is fundamentally superior to Fisher's original technique is of little concern because both are contrary to observation oriented modeling. One could argue, however, that the sophisticated hybrid has only exacerbated the misunderstandings held by researchers regarding NHST. Because of its complexity, fundamental features of its probabilistic nature, such as random sampling and *in the long run*, are also commonly misunderstood and routinely overlooked in modern treatments of NHST.

This page intentionally left blank

CHAPTER 6

Modeling and Inferential Statistics

Contents

Introduction 87
Modeling in Science 88
The Bystander Effect Model 94
Estimating Population Parameters in NHST 99
Modeling Persons, Testing Aggregates 104
Observation Oriented Modeling 108

INTRODUCTION

A paradigm, according to Thomas Kuhn (1970), is a complex of accepted philosophical premises, established methods of investigation, and reigning theoretical models that guides scientific investigation. Kuhn argued that science progresses through continual "revolutions" in which old paradigms slowly decay from within and then fall when suitable incommensurable paradigms arise to take their place. As intriguing as this thesis has proven to be throughout the years, it has not been without its critics, most notably Karl Popper.[1] It also stands in contrast to the research of Pierre Duhem discussed in Chapter 5, which showed that scientific progress, even through the so-called Dark Ages, may not be predicated on the notion of incommensurable paradigms. On the other hand, Kuhn's thesis is consistent with the notion that science is an embodied endeavor, which is to say simply that it is human. Scientists are not sensation sorting machines as Pearson would have us believe, and much of what drives their research lies beyond the realm of logic or the ostensibly objective search for knowledge regarding how things work in nature. Like experts in any other occupation, scientists rely on their illative sense to plan, execute, and interpret their experiments. Moreover, irrationality, egoism, peer pressure, and politics may all enter into the scientist's day-to-day activities as well as his or her theoretical reasoning. The notion of paradigm therefore

[1] For criticisms of Kuhn's thesis, see the papers in Lakatos and Musgrave's *Criticism and the Growth of Knowledge* (1979). Karl Popper's chapter is titled "Normal Science and Its Dangers" (pp. 52–58).

Observation Oriented Modeling
ISBN 978-0-12-385194-9, Doi:10.1016/B978-0-12-385194-9.10006-4

© 2011 Elsevier Inc.
All rights reserved.

serves as a useful reminder of the embodied nature of science, and meditation on this idea should lead to a critical evaluation of the philosophical premises and methodological practices that guide any scientist's research.

The critique of the philosophy and methodology of the Pearsonian–Fisherian approach begun in Chapter 5 is extended to the pragmatic realm of scientific modeling in this chapter. Although considerations of paradigms may in the final analysis be superfluous (replaceable with embodiment), the importance of models to scientific investigation cannot be gainsaid. As noted previously, Kuhn considered models to be subordinate to paradigms, but the argument could certainly be made that his book *The Structure of Scientific Revolutions* (1970) chronicles the unpredictable but steady growth of scientific models in physics, chemistry, and biology over the course of history. Rather than viewing scientific progress through the lens of a Marxist-like scheme of paradigm shifts, then, focus can instead be placed on the continual development of models that offer ever-increasing levels of depth and complexity as explanations of ordered nature.[2] From this vantage point, it is argued in this chapter that the Pearsonian–Fisherian approach actually thwarts scientific progress in psychology by retarding the development and growth of its models. In comparison, it is shown that observation oriented modeling, which rests upon the foundation of philosophical realism, restores modeling to its rightful place at the center of scientific investigation. These arguments are made concrete through the re-examination and re-analysis of a popular study published in the 1960s.

MODELING IN SCIENCE

Perusal of any peer-reviewed journal of psychological research published within the past 100 years yields a dizzying array of theoretical models, many of which are presented and explained solely through the use of language without any recourse to mathematics, symbols of logic, or visual schematics.

[2] The description "Marxist-like" is used cautiously here as an analogy, although some may construe it as more descriptive or explanatory. See Cohen (1985), particularly Chapter 25. Clifford Connor's *A People's History of Science* (2005) also draws on Marxist themes and serves as a reminder that scientists are necessarily enmeshed in their respective cultures. Connor particularly and worthily explores the importance of artisans upon the work of different scientists, but various speculations throughout his text must be carefully weighed against the work of other historians.

For example, when describing his model of emotional change in psychotherapy, Pascual-Leone (2009) explains,

> *Through prolonged exposure to the fearful stimulus [in therapy], the initial intense affective reaction is gradually attenuated through habituation among other processes. Moreover, with repeated exposures there is a reduction of the initial starting level of distress across each session. The effect is a decrease in distress both within each session and across treatment. This describes effective emotional processing in the treatment of anxiety as occurring in a pattern … described as sawtoothed—a "two-steps-forward, one-step-back" progression during therapy.*
>
> ***(p. 114)***

Other presentations of models employ visual schematics to represent the relationships between variables, such as the path analysis model in Figure 6.1. In this model, the researcher posits that variation in women's decisions (labeled "intention" in the model) to obtain prenatal screening tests for Down's syndrome is a function of variability in a number of subjective judgments and attitudes (van den Berg et al., 2008). Still other schematic models represent structural features of phenomena

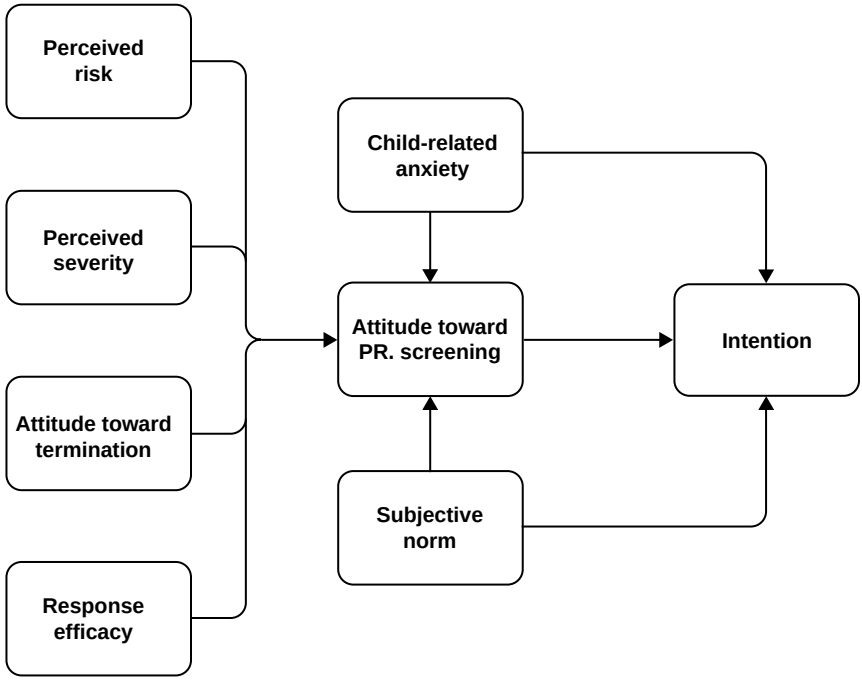

Figure 6.1 Path model depicting a mother's choice to obtain prenatal screening for Down's syndrome

that may not readily be considered as composed of both matter and form, such as Atkinson and Shiffrin's (1968) model of memory shown in Figure 6.2.

Finally, some models in psychology that are relatively infrequent are presented primarily in mathematical and logical form. Vladimir Lefebvre's algebraic model of self-reflexion is one such model, which can be expressed as follows:

$$\begin{array}{l} \quad a_3 \cdots \star \cdots b_3 \qquad a_4 \cdots \star \cdots b_4 \\ a_2 \cdots\cdots\cdots \star \cdots\cdots\cdots b_2 \\ A_1 = a_1 \end{array} \qquad [6.1]$$

This model is meant to describe how individuals integrate input from the environment (a_1) and various subjective states regarding positive or negative views of one's self (a_2, a_3, and a_4) and another person (b_2, b_3, and b_4) to "compute" a response (A_1) to a binary decision task, such as responding "agree" or "disagree" to an item on a questionnaire. Although on the surface this model appears to be a power function based on products, the equation is actually solved using a type of Boolean algebra referred to as gamma algebra (Lefebvre, 2001). The exact details of the model are of little concern here because it is presented solely to exemplify how psychological phenomena can be represented using mathematics.

As with all things scientific, the importance of models can be traced back to the time of ancient Greece—for example, in Aristotle's (384–222 BC) sketch of the rainbow in book III of his *Meteorology*, in Hippocrates' (460–370 BC) physiological model of the four humors, and in Heraclitus' (535–475 BC) early view of the heavenly bodies revolving around some

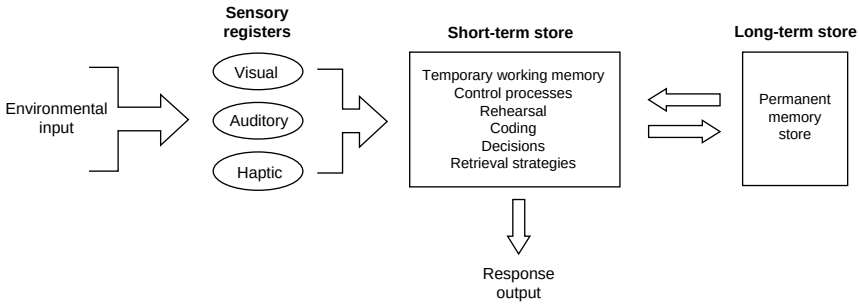

Figure 6.2 Classic model of human memory

vague "central fire." In the Middle Ages, Albertus Magnus (1206–1280) contemplated a spherical model of the earth and echoed the importance of experimentation, "Experimentum solum certificat in talibus," embodied in the ancient work of Archimedes (287–212 BC) and later outlined by Robert Grosseteste (1168–1253) and Peter of Spain (1215–1277). Two centuries later, William Gilbert (1540–1603) thoroughly modeled the loadstone (magnet), and then, in what is the most notorious example of one model pitted against another, Galileo Galilei (1564–1642) successfully overturned the Ptolemaic model of the solar system in the early 17th century. Finally, the revolutionary success of Gottfried Leibniz (1646–1716) and Isaac Newton's (1643–1727) explanations of matter and motion solidified the importance of formal mathematics as an invaluable tool in the modeling of nature.[3]

These few examples do little justice to the innumerable and remarkable scientific models that have been created since the time of ancient Greece. They also do not make clear the dynamic nature of models because, as philosophers of science are apt to point out, scientific knowledge is cumulative, building systematically—if not always predictably—upon itself over time. The accretion of scientific knowledge is necessarily wedded to the growth and development of its models.[4] One can consider the early model of the atom and its extraordinary evolution over the years or the development of the model of cellular biochemical pathways.[5] The very appearance of these models, as they are normally presented schematically, reveals the structures (formal causes) and time-ordered steps (efficient causes) that underlie the transformation of matter (material causes) toward diverse, natural end states (final causes). Their historical development also reveals an important Aristotelian concept of working from what is better

[3] The reference to Heraclitus is taken from Boynton (1948, p. 2). Albertus Magnus is quoted ("experiment is the only safe guide in such investigations") by Kennedy (1907). A. C. Crombie (1961) discusses Robert Grosseteste and Petrus Hispanus (Peter of Spain). William Gilbert's (1600/1952) complete work on the loadstone can readily be found in Encyclopedia Britannica's *Great Books of the Western World* series. Finally, the great minds that shaped Western theological, philosophical, and scientific thought are presented in readable fashion in Watson and Evans' excellent textbook, *The Great Psychologists* (1991).

[4] Harré (1970) offers an excellent discussion of the different forms of models employed by scientists and discusses their importance as well.

[5] William Wallace (1996) provides a readable overview of the history of the atomic model. A copy of the cellular biochemical model can be viewed online at http://www.expasy.org/cgi-bin/show_thumbnails.pl?2. A copy can be purchased from Roche Applied Science.

known to what is lesser known in the process of scientific modeling. Once the original model of the atom, which has the appearance of a solar system with electrons revolving round a nucleus of protons and neutrons, was well-established via formal mathematical analysis and experimentation, scientists began to work out the details of the components of the model and their interrelationships—for example, adding directional spin to the electrons, determining elliptical orbits, expanding the dimensionality of the model, and differentiating among forces within the atomic nucleus. The model of cellular biochemical pathways underwent a similar evolution and is quite remarkable in its integrated complexity.[6]

When such models are coupled with the scientific method, both the heart and the soul of *scientia naturalis* are truly revealed. This science of nature was discussed in Chapter 2, and as *scientia* it is knowledge that is demonstrable through reasoned argument. As *naturalis* it is verified primarily through the senses. In other words, the modern scientist's reasoning is embodied or expressed in his or her model, and its content and structure are tested via controlled observations and experimentation. Thomas Aquinas spoke of the early scientific models of his teacher, Albertus Magnus, and his contemporaries (e.g., Roger Bacon and Robert Grosseteste) as intermediary sciences because they stood between the realms of pure abstract mathematical reasoning and the concrete world (Aquinas, 1963). Mathematics itself is devoid of individuals. For example, the equation $2 + 2 = 4$ is not wedded to particular things or events in the universe; instead, it can be applied to any countable things and events in the universe. Yet in a person's day-to-day experiences, he or she encounters only individual people, places, and events. Models provide a bridge between this concrete realm of lived experience and the realm of concepts, such as mathematics and geometry. The north and south lanes on this metaphorical bridge, however, are not equal—meaning that, consistent with Aristotle, the overall direction of movement in scientific knowledge must be from what is more known in the observable to what is less known. Through the lens of Jacques Maritain's (1951) philosophy of nature, a scientific model is seen as first derived

[6] In his book *Wretched Aristotle*, Jude Dougherty (2009) recounts (see pp. 69–72) the work of Lise Meitner and Otto Frisch, who relied on a metaphorical model of a drop of water dividing into two droplets as their inspiration for their discovery of nuclear fission. This is an excellent example of what is essentially a crude model that would lead to particular mathematical computations, controlled experiments, and a deeper understanding of nature.

through sense experiences of the world. The model is then turned back toward nature in an attempt to descend progressively deeper into the observable world:

> *We have said that empiriological analysis resolves the object into the observable or measurable elements and thus goes from the observable to the observable, always remaining on the level of sensible operations, of obtainable observations or measurements.*
>
> ***(p. 77)***[7]

The scientist's model is therefore circumscribed, or constrained, by various features, such as the observable nature of it components or its capacity for generating sensible operations (e.g., measurement) that are necessary to evaluate its content and form. In other words, a scientific model must generate hypotheses that are at least in principle testable by empirical means—that is, certain physical operations performed under controlled conditions.

The philosopher, by contrast, is also interested in comprehending reality but starts from the senses and ascends to the intelligibility of nature. A modern scientist studying the atom, for instance, must work with mathematics and use sophisticated tools (e.g., a supercollider) to gather observations of particles or forces included in the model. In this way, the atomic model can be modified and enhanced to provide a richer understanding of nature. The philosopher is not tied to the supercollider or even to mathematics, and he or she can instead ask questions about the intelligible essence of atoms. What is the essence of a given atom, and what does it mean to speak of "essence" in the first place? What does it mean to measure some thing in the universe or to posit that it exists quantitatively rather than qualitatively? Questions such as these are of equal importance to those asked during an experiment conducted with a supercollider, but they are oriented toward intelligible being rather than the observable and measurable.[8] In a sense, the philosopher addresses Einstein's famous enigma, "The most incomprehensible thing about the world is that it is comprehensible," by working out why and how the universe is intelligible. As explained by Pope Benedict XVI (2006) in his address to the faculty at the University of Regensberg,

[7] Citing Maritain here is not to be construed as endorsement of his entire philosophy, which some have argued may contradict a truly Thomistic understanding of science (see Nogar, 1951).

[8] Anthony Rizzi (2004) demonstrates how the realist philosophy of Aristotle and Aquinas can be used to understand many apparent paradoxes and quandaries of modern physics.

> *Modern scientific reason with its intrinsically Platonic element bears within itself a question which points beyond itself and beyond the possibilities of its methodology. Modern scientific reason quite simply has to accept the rational structure of matter and the correspondence between our spirit and the prevailing rational structures of nature as a given, on which its methodology has to be based. Yet the question why this has to be so is a real question, and one which has to be remanded by the natural sciences to other modes and planes of thought—to philosophy and theology.*

Donning the hat of a scientist rather than a philosopher, the research psychologist interested in employing the models at the beginning of this chapter recognizes that they all share a common goal, namely to represent an observable natural system that is assumed to be a lawful organization (final cause) of discrete structures (formal and material causes) and time-dependent processes (efficient causes). The scientist will also recognize that the growth of *scientia naturalis* will be limited by the development of these models. Stated differently, psychologists' understanding of the natural processes they are studying (i.e., dynamic, living persons) will depend on their ability to create, develop, and accurately test models such as those presented previously. It follows that philosophical premises, laboratory practices, or methods of data analysis that thwart model development are necessarily thwarting the advancement of scientific knowledge as well. Consistent with the criticisms of null hypothesis significance testing (NHST) in the previous chapter, how this method of data analysis has, for the most part, played the role of spoiler in the development of successful psychological models can now be explained. Specifically, this unfortunate state of affairs is due to (1) persistent confusion regarding the essence of NHST as the estimation of population parameters and (2) a failure to reconcile the logical inconsistency between modeling at the level of the aggregate versus modeling at the level of the observation. Once these points are presented and clarified, an alternative path for testing models can be cleared.

THE BYSTANDER EFFECT MODEL

In his book *Forty Studies That Changed Psychology*, Hock (2009) includes the 1968 study by John Darley and Bibb Latané, which was published on the heels of a tragic event that took place in New York City. Specifically, a young woman had been stabbed to death on a residential street while 38 people witnessed the attack and failed to offer any type of help. Not a single person even placed a phone call to the police. Adding to the shock

of the story, the attack was intermittent, lasting up to 30 minutes before the young woman was finally murdered. Although the accuracy of the original accounts of this incident have been seriously questioned, Darley and Latané formulated several explanations of the bystanders' alleged behavior (Manning, Levine, & Collins, 2007). They then created a situation in their laboratory in which undergraduate students witnessed another individual in distress and had to decide to offer help or sit idly. In what has been dubbed the "bystander effect," the results of their study purportedly showed that the students were not likely to help individuals in distress when others were known to be witnessing the same events. The design of Darley and Latané's study was so clever, and the results were so interesting, that it is easy to understand Hock's decision to include it in his book. For the current purposes, it is instructive to describe the study in detail as well as Darley and Latané's explanations of the bystander effect.

Imagine you are a participant in Darley and Latané's original study. At the appointed time, you arrive at the designated building and room to find a male experimenter waiting in an area connected to a hallway that leads to other rooms. The experimenter thanks you for coming and then escorts you down the hallway to a private room equipped with a communication system. He explains that other participants are waiting in adjacent rooms and that you will all be discussing the difficulties associated with college life, which is the purpose of the investigation. Given the personal nature of what might be discussed, he states that the intercom system is a useful way to protect your anonymity and to reduce possible embarrassment when sharing personal information. Finally, the experimenter tells you that he will not be listening in on the conversation, and he describes that the system is set up for one-way communication. It will automatically switch between discussants after a set period of time. You are therefore to wait your turn, at which time you can contribute to the discussion, and then the system will automatically switch to the next person. The experimenter then leaves, closes the door, and the discussions begin. After the first round in which everyone has spoken, one of the individuals begins speaking but soon grows distressed, then stammers, refers confusedly to seizures, and finally becomes silent. You cannot ask how the person is doing because of the one-way intercom, and you are not certain if the person fell silent because of some personal distress or because the intercom switched to the next person. What do you do? Do you sit and wait, or do you exit the room and look for help? If you choose to wait, the experimenter allows you to sit in the room for up to 6 minutes. When you finally meet with the experimenter again, he proceeds to tell you

that the person suffering the apparent seizure was a recorded confederate (fellow experimenter) who was only acting. The experimenter then asks you to complete a checklist reporting your feelings and thoughts and to fill out a number of questionnaires assessing traits such as Machiavellianism and authoritarianism. After completing these tasks, you are dismissed.

Whether or not participants sought help before the seizure victim was cut off by the automated intercom system was one of the primary dependent variables (i.e., outcome variables) of Darley and Latané's study. The other primary dependent variable, which will not be considered herein, was the time measured in seconds between the beginning of the distressed person's seizure and the participant's exiting the room to seek help. What was the independent variable in the study? Unbeknownst to each student—unless the experimenter accidentally tipped his hand—there were no other participants in the study. All of the other voices heard over the intercom were also recorded and played back. Moreover, depending on the condition to which each participant was assigned, he or she heard one other student, two other students, or four other students in the discussion. These different conditions constituted the independent variable of the study and are referred to herein as the 0-bystander, 1-bystander, and 4-bystander conditions, respectively. A participant in the 0-bystander condition spoke with one other person who, after one turn exchanging thoughts, suffered from an apparent epileptic seizure. Therefore, there were no bystanders because the participant was the only person aware of the other's difficulties. A participant in the 1-bystander condition heard two other people—the seizure victim and one other discussant (bystander). Finally, in the 4-bystander condition, the participant was led to believe that five other people were speaking—the seizure victim and four other discussants.

Darley and Latané posited that students in the conditions with one or more bystanders (particularly the 4-bystander group) would respond more slowly to the emergency and would not seek any form of help compared to the participants in the condition with no bystanders. They offered three efficient cause explanations for the anticipated results:

1. The responsibility for helping the victim would be diffused among the bystanders in the 1- or 4-bystander conditions, thus resulting in inaction from the participant.
2. The blame for failing to take action would be diffused among the bystanders, again resulting in inaction from the participant.
3. The participants in the two bystander conditions would fail to take action because they would assume that someone would help the victim.

A schematic model of the 0- and 4-bystander conditions representing the third causal hypothesis is presented in Figure 6.3. The 1-bystander condition is omitted to streamline the analysis and discussion of the model. The model is meant to represent a short span of time (~1 minute) during the experiment beginning at the moment the first person talks over the intercom. It shows the distressed person (DP), the confederate bystanders (B), and the participant (P). It also shows that the entire person must be considered in the model, including the person's sensory connection to the world and his or her self-reflective consciousness. In general, psychological models will strive to integrate the embodied (senses, physiology, and biochemistry) acts (behaviors), powers (abilities and capabilities to achieve different ends), and habits (dispositions as well as consistent behaviors) of a person. Models such as the one shown in Figure 6.3 are thus referred to as integrated models in observation oriented modeling. Another feature of the current model is that none of its components are structured as continuous quantities; rather, the different components are specified as categories. For example, the person either perceives or fails to perceive the distressed confederate, and the person either judges the confederate to be in distress or fails to make this judgment. The specific observations that will be analyzed are also categorical, ordered as 0 or 4 (0 or 4 bystanders) and as 0 or 1 (seeking help or failing to seek help). As a nonquantitative model, logical statements will be most appropriate for connecting its various components. The importance of the quantitative/qualitative distinction as it pertains to psychology is discussed in more detail in Chapter 8.

To be sure, the current integrated model is simplistic. Its components could be clarified or expanded, or additional components, quantitative or qualitative, could be added. For instance, the process by which a person actually reaches a conclusion in his or her intellect to help or remain seated can be clarified. Is it a complex judgment task or a simple one? Does the participant incur a sense of guilt in the process? Do the two other causal mechanisms (diffusion of responsibility and shifting blame) posited by Darley and Latané come into play here? How much emotionality is involved in the process? Ostensibly, Lefebvre's model (see Eq. 6.1) could be employed to formally express this component of the model. In terms of expanding the model, it may be that some forms of witnessed distress, such as a rape or physical attack, are more frightening or perhaps less ambiguous than others. A participant who has helped a loved one with seizures in the past may also be expected to react differently to the situation than one who has no experience or knowledge of seizures. Pursuing any of these modifications

if 4* then
1
P judges DP to be in authentic distress
AND judges that E is unaware of distress
AND at least one B will obtain help
if 0* then
0
P judges DP to be in authentic distress
AND judges that E is unaware of distress
* AND all perceptions are brought into conscious awareness
Distressed Person (DP)
P perceives DP
via intercom
P perceives communication
between B and DP when
4 bystanders present
P perceives B
via intercom
Participant (P)
0 or 4
Bystanders (B)
Experimenter (E)
if 0 then P
seeks E for help
ELSE P stays in
room

Figure 6.3 Integrated model of the 0- and 4-bystander conditions

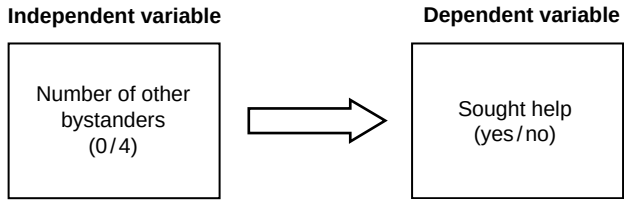

Figure 6.4 Simple variable-based model of the 0- and 4-bystander conditions

would exemplify the notion of scientific progress discussed previously. The researchers would continually develop their model, moving from what is well known about reality to what is lesser known. Consequently, their model would provide an increasingly richer explanation of nature via carefully controlled observations.

How then, did Darley and Latané test their first model and then proceed toward its development? Unfortunately, they inherited and followed the Pearsonian–Fisherian vantage point that had already become deeply entrenched in psychology by the 1960s. This resulted in a number of fundamental miscues in their approach toward modeling and, by necessity, their scientific research. The foremost misstep is that Darley and Latané did not spell out their model as an organization of discrete structures and time-dependent processes such as those seen in Figure 6.3. Instead, as noted previously, they conceptualized their study primarily in terms of an independent variable abstractly connected to a dependent variable, as shown in Figure 6.4. The origins of this simple model can be traced back to Fisher's early work in which he considered a variable to be a quantitative dimension of nature that could assume different values; for example, pea plants can vary in their individual heights. Fisher created his analysis of variance technique to divide the observed variation of a dependent variable into separate parts: systematic variation attributable to the independent variable and nonsystematic variation with no specified source. This technique furthermore entailed representing systematic variation as estimated population parameters, such as the estimated mean difference in height between two generations of pea plants.

ESTIMATING POPULATION PARAMETERS IN NHST

It is toward the task of estimating population parameters that Darley and Latané turned in the analyses of their data. Considering only the results for the 0-bystander and 4-bystander conditions in Table 6.1, it can be seen that

Table 6.1 Observed results for 0- and 4-bystander conditions of Darley and Latané's experiment

		Sought help	
		Yes	**No**
No. of bystanders	0	11	2
	4	4	9

participants in the 0-bystander condition sought help with greater frequency than those in the 4-bystander condition. This result can be expressed in terms of an odds ratio computed as (11/2) ÷ (4/9) = 12.38. Following NHST, this odds ratio can be considered as an estimate of some fixed population parameter.[9] To begin, the student participants are assumed to represent a larger population of students who could be assigned to the two conditions and who would yield a population odds ratio of exactly 1. This assumption is tantamount to assuming that the null hypothesis, which posits no relationship between the independent and dependent variables in the population of assigned students, is true. The observed odds ratio can then be tested for approximate statistical significance on the basis of an assumed continuous normal distribution using a well-known formula:

$$z = \frac{\ln(OR)}{\sqrt{\frac{1}{x_{11}}+\frac{1}{x_{12}}+\frac{1}{x_{21}}+\frac{1}{x_{22}}}} = \frac{\ln(12.38)}{\sqrt{\frac{1}{11}+\frac{1}{2}+\frac{1}{4}+\frac{1}{9}}} = \frac{2.52}{.95} = 2.65. \quad [6.2]$$

OR is the observed odds ratio, and the x_{11}, x_{12}, etc. values are the frequencies from Table 6.1. The result is statistically significant when compared to the conventional p value criterion of .05 and corresponding two-tailed z critical value of 1.96; consequently, the relationship between the independent variable (number of bystanders, 0 or 4) and the dependent variable (seeking help, yes or no) is judged to be greater than a 1:1 odds ratio in the population. Based on the data, the population value is estimated to be 12.38, meaning that students in the 0-bystander population will seek help at a rate 12.38 times greater than will students in the 4-bystander population.

[9] Darley and Latané estimated a population chi-square value in their statistical analyses rather than an odds ratio. An odds ratio is used here because it is a user-friendly and common metric.

NHST is a probabilistic scheme that furthermore recognizes the computed odds ratio as one of an infinite number of point estimates of a fixed population value. It is an estimate because it is based on a sample of observations rather than the entire population, and different odds ratios would be obtained from independent samples drawn from the same population. Any or none of the sample odds ratios could equal the population value. It is therefore prudent that a single point estimate of a population parameter be accompanied by a confidence interval. For the current odds ratio (12.38) and critical p value (.05), the 95% confidence interval is

$$\begin{aligned} &\ln(OR) \pm (z_{critical})\left(\sqrt{\frac{1}{x_{11}}+\frac{1}{x_{12}}+\frac{1}{x_{21}}+\frac{1}{x_{21}}}\right) \\ &= \ln(12.38) \pm (1.96)\left(\sqrt{\frac{1}{11}+\frac{1}{2}+\frac{1}{4}+\frac{1}{9}}\right) = 2 \pm 4.44 \end{aligned} \quad [6.3]$$

The exponent of each value from the equation is taken to yield the lower and upper bounds of the confidence interval. Accepting all of the assumptions of NHST, Darley and Latané could hence conclude with 95% confidence that the interval 1.83 to 83.77 includes the true population odds ratio. The results of their study thus supported a much smaller population association between the independent and dependent variables (1.83) as well as an extraordinarily larger association (83.77; the students in the 0-bystander condition seek help with a frequency ~84 times greater than do the students in the 4-bystander condition). This wide interval clearly indicates that the estimated magnitude of the relationship between the independent and dependent variables in the population is highly imprecise. Following the conventional reporting practices of their time, Darley and Latané unfortunately did not compute a single confidence interval for their analyses. Had they computed confidence intervals, they may have qualified their interpretations of the results and then sought to replicate their study employing a larger sample size to yield a more precise confidence interval. Like their peers and the psychologists of today, however, they did not understand the essence of NHST as the estimation of population parameters. Instead, it is likely Darley and Latané viewed their statistically significant results as a suitable substitute for conducting an exact replication. Although they made no explicit claim to this belief, past surveys have shown that sizable proportions of individuals who employ and teach NHST believe that a statistically significant result with a p value equal to .05 indicates the study will successfully replicate 95 out of 100 trials. This belief has long been

shown to be fallacious, and its persistence is puzzling given several well-known attempts at correction in the mainstream psychological literature.[10,11]

Having traded exact replication for NHST, Darley and Latané next attempted to establish the link between two abstract variables ("number of other bystanders" and "sought help") using a wide variety of experimental methods and procedures. They dubbed this relationship "bystander intervention" and in one study blew smoke through a vent while students completed a packet of questionnaires alone or in the presence of two confederates or two other participants. In another study, they staged a theft in a convenience store while the clerk was busy in the back of the store. Participants witnessed the theft while alone in the store or in the presence of another customer. In yet another study, they recorded whether or not participants would help a marketing researcher in an adjacent room who had apparently suffered a mild injury while retrieving a book from a high shelf. The independent variable in each of these studies was the number of

10 The most visible effort to correct the misunderstanding and misreporting of statistical results came from a task force appointed by the American Psychological Association that reported their conclusions in the *American Psychologist*, which is the most widely circulated social science journal in the world (Wilkinson et al., 1999). Gigerenzer (2004) reviews a number of other efforts, including attempts by editors to enforce policies regarding the reporting and interpretation of probability values computed from NHST. He concludes that such efforts have been largely fruitless.

11 It is important to address several recent attempts to connect the p values from NHST to the probability of an exact replication. First, Killeen (2005) derived a probability value, p_{rep}, which purportedly represented the probability that a particular statistical result would replicate in a random sample. Unfortunately, errors were discovered in his derivation, and it was later shown that p_{rep} is essentially an index of statistical power (i.e., the probability that an estimated aggregate effect will be declared as statistically significant in an assumed randomly and independently drawn sample of observations) (Maraun & Grabriel, 2010). Second, Greenwald, Gonzales, Harris, and Guthrie (1996) argued that Fisher's original p value is monotonically related to the probability that an effect will replicate if one defines replication as "rejecting the null hypothesis and concluding the same direction of effect" in the context of NHST. They are careful to point out that this is not the standard definition of replication, which they see as estimating the difference between two point estimates of parameters. From the perspective of observation oriented modeling, we would challenge their standard definition of replication as well, given that it potentially involves the estimation of population parameters in a random sampling scheme. Indeed, this allegedly more standard form of replication is likely to be foreign to most practicing scientists from a wide variety of domains. To stave off confusion, then, it is best to avoid the term "replication" altogether when describing Greenwald et al.'s interesting article. Their arguments are most parsimoniously described as showing that Fisher's p value in NHST is monotonically related to statistical power.

bystanders, and the primary dependent variable recorded whether or not participants intervened in the emergency (e.g., helping the person in distress or reporting the theft). Darley and Latané considered each of these studies to support the so-called bystander effect and to illuminate their theoretical model of the intervention process. In their 1970 book *The Unresponsive Bystander: Why Doesn't He Help*? they claimed to have been testing an integrated model like that presented in Figure 6.3:

> *We suggested that before a bystander will intervene in an emergency, he must notice that something is happening, interpret that event as an emergency, and decide that he has a personal responsibility for coping with it. If a person fails to take any of these steps in a direction favoring intervention, he will fail to intervene.*
>
> ***(p. 121)***

This claim, however, is dubious for the simple fact that the model did not appear in verbal or schematic form in their earliest published studies. It instead appears in the 1970 book as a simple heuristic for organizing the results of 12 different studies, none of which were exact replications of previous studies. It is therefore, at best, a vague model that was not used to guide research in specific ways. Its connection to the particular studies is also uncertain. For instance, Darley and Latané suggested the study involving the marketing researcher retrieving a book and the study with the smoke-filled room should be viewed together as providing evidence for the first step in the decision process described previously, namely the person noticing that something is happening. The primary independent variable in both studies, however, was the number of bystanders, which was correlated with helping behavior. Indeed, the studies were designed on the basis of the simple model in Figure 6.4, not the verbally described model in their 1970 book. The centrality of the number of bystanders was admitted by Darley and Latané when they wrote, "We suggested that individuals might be less likely to notice an emergency when they are in the presence of others, since, in public, it is embarrassing to be too attentive to others" (p. 87). Moreover, they noted that every participant in each of their studies had noticed that some event requiring action had taken place, although some participants noticed the event sooner than others:

> *Subjects in the smoke-filling room were slower to notice the smoke when there were others present, but this effect was rather small, significant only statistically. Subjects were not slower or less likely to notice the theft of money when others were present. All subjects in the injured woman study seemed to notice her accident immediately.*
>
> ***(p. 87)***

If they had truly been working from an integrated model, Darley and Latané would have recognized that the first necessary step in their model was not established experimentally, and that their program of research should have followed an entirely different path.

NHST is wedded to simple variable-based models such as the one shown in Figure 6.4, and its essence is the estimation of population parameters under various assumptions. By adopting this approach, it was perhaps inevitable that Darley and Latané would be led down the wrong lane of the metaphorical modeling bridge discussed previously. After their initial study, they made a move toward further abstraction in which they prematurely divorced their simple variable-based model from its original methods and procedures. Recalling the words of Maritain, "empiriological analysis … goes from the observable to the observable," it is clear that Darley and Latané moved from the observable aspects of their study to an abstract model of bystander intervention that was considered to be largely independent of any given study's methods and procedures. If they had sketched and elaborated an integrated model of nature like the one in Figure 6.3, they would have systematically investigated its discrete structures and time-dependent processes. They would also have endeavored to clearly distinguish between their three causal hypotheses (diffusion of responsibility, shifting blame, and assuming help), thus generating a truly organized body of knowledge of nature through its causes. Moreover, if they had sketched integrated models for each of their studies, Darley and Latané would likely have realized that the variable models they were testing were in fact incommensurable and incapable of yielding the steady accretion of knowledge so characteristic of *scientia naturalis*. Given the Pearsonian–Fisherian approach of the time (and confusion regarding its application), however, they were led to disregard the importance of exact replications and to enter a cycle of estimating population parameters for abstract, simple models that could hardly be used to descend progressively deeper into the observable world. In the end, they were left with the task of cobbling together a number of studies that were only superficially related. Unfortunately, this characterizes the vast majority of psychological research published since the ascendency of the Pearsonian–Fisherian approach.

MODELING PERSONS, TESTING AGGREGATES

Darley and Latané's statistical analyses for their original bystander study were not conducted on the basis of the integrated model shown in Figure 6.3 but

instead the simple variable-based model shown in Figure 6.4. Even for this latter model, however, the estimated population parameter was only loosely related to the conceptual and pragmatic goals of their classic study. Darley and Latané failed to realize this point, given their adherence to the Pearsonian–Fisherian tradition. Recall that they offered three explanations for the bystander effect, and that each of these explanations involved the cognitive processes (e.g., shifting blame to other bystanders) of a person failing to help another in distress. Their explanations were therefore presented at the level of the persons in their study. They did not attempt, via theory, to predict proportions pertaining to the frequencies with which participants would seek help in the different bystander conditions. This would turn their theory into a model of aggregates (i.e., responses collapsed or averaged over individuals) rather than a model of persons and would require replacing the single arrow in Figure 6.4 with a predicted proportion of persons seeking help for the seizure victim in each of the two conditions. Alternatively, they could predict a population odds ratio for the model. Making such predictions, however, runs contrary to the mechanics of NHST, which are based on the assumption of the null hypothesis (an odds ratio equal to 1:1 in the population in this case) being true. Given this assumption, the population parameter is then estimated post hoc from the data and evaluated for statistical significance. Significant changes would also be required for the integrated model in Figure 6.3 if it were to represent frequencies. For instance, the qualitative links representing the 0-bystander person's judgments regarding the distressed person and bystanders (0, 4 bystanders; help/no help) and the experimenter (if no bystanders, then participant seeks help, else participant stays in room) would have to be altered substantially to entail predicted frequencies. The processes underlying these frequencies could furthermore be developed to explain exactly why such judgments are made by some individuals and not others. As it is currently drawn, the model predicts that every person in the 0-bystander condition will help the person in distress. Alternatively, the frequencies could be represented as *a priori* probabilities that pertain to repeated observations made on a single individual. This strategy would keep the model at the level of the person, and it would mimic how quantum physicists employ probabilities in highly specified roles in modern models of individual atoms—not aggregates of atoms (Gigerenzer et al., 1989, Chapter 5).

The centrality of estimating aggregate population parameters in the Pearsonian–Fisherian tradition also prevented Darley and Latané from asking important questions about their observations. An obvious question

that should have arisen from the bystander data presented previously is, Why did two people in the 0-bystander condition fail to seek timely help for the seizure victim even though they were the only witnesses? It was only after the seizure victim fell silent and it was their turn to talk did these two participants finally emerge from the room. Similarly, in the 4-bystander condition, four people sought help quickly even though other bystanders were present. Why? These questions were never asked because the focus was on estimating population parameters. The individual cases were important only to the extent that they were used to compute a frequency or some other aggregate statistic (e.g., mean, median, and standard deviation) that could be used in NHST. This bias prevented Darley and Latané from connecting different observations made on the students in their study. Most notably, 20 participants claimed, at the conclusion of the study, to have guessed the seizure was some sort of fake. Was this claim made by the two participants who were slow to report the seizure in the 0-bystander condition and by the four people who sought help quickly in the 4-bystander condition? Although NHST could be used to assess differences in frequencies between groups with regard to guessing the seizure was fake, it simply does not provide the framework for answering this question because it cannot be used to connect individual observations in a flexible and meaningful way.

The sacrifice of anomalous results, as well as deviant cases, to an estimated aggregate population parameter is disconcerting on yet a deeper level. Philosophers and historians of science have long recognized the importance of such results or cases in the advancement of scientific knowledge. With regard to anomalous cases, Thomas Kuhn devoted an entire chapter of his influential book *The Structure of Scientific Revolutions* (1970) to the importance of serendipity. In recounting the discovery of X-rays, he described how the physicist Roentgen noticed that a barium platinocyanide screen glowed unexpectedly during an experiment involving cathode rays (see Kuhn, 1970, p. 57). Rather than ignore this anomaly, Roentgen stopped his experiments and spent 4 weeks convincing himself that the effect was not at all due to cathode rays but, rather, to some other agent similar to light. This agent would later be known as an X-ray. In the realm of NHST, in which the focus is on estimating population parameters, often with no empirical basis, there is very little room for such serendipitous findings. Results are bracketed into binary categories as significant or nonsignificant, and unusual results can be attributed to the dumb luck of occasionally encountering strange outcomes in a sampling process.

NHST also ignores the potential importance of deviant cases, or "outliers" as they are commonly called. These outliers are considered as nuisances because they may disproportionately influence the estimate of a population parameter—for instance, when an extreme outlier greatly influences the mean computed from a small sample of observations. Researchers are trained to cope with outliers using a number of strategies, such as case deletion, data transformation, Windsorization, and switching to a nonparametric statistic for analysis. All of these techniques are intended to preserve the validity of the estimated population parameter, and they work against the scientist's natural disposition to learn from unusual observations. Indeed, scientists can make effective use of outliers in at least two ways. First, when developing a measurement technique, outliers often provide the clearest information regarding how the technique can be improved to yield more reliable observations. Consider a father and son building a simple catapult to demonstrate the relationship between potential and kinetic energy. The spring device and release mechanism must work consistently for the relationship to emerge clearly. Once a moderate degree of consistency is achieved, it is from the outliers that the father and son will learn most clearly how to further adjust the spring and release mechanism. Second, outliers may provide clues for modifying one's model. If Darley and Latané had used the integrated model in Figure 6.3 and paid close attention to the deviant cases in their study, they may have found ways to improve the model. Again considering the two people who failed to seek help quickly in the 0-bystander condition, perhaps these students were disposed to inaction when seized by extreme fear. If this were the case, then the component of the model involving the conscious decision to seek help could be expanded and directly tied to emotional states. By adopting the simple model in Figure 6.4 and NHST, however, Darley and Latané put themselves in the unfortunate position of simply ignoring the cases that deviated from their expectations.

In summarizing Darley and Latané's original bystander study, it is as if they began implicitly with several detailed models such as the one shown in Figure 6.3 to describe the inter- and intrapersonal processes that would help them to understand not only the Kitty Genovese murder but also all individual acts of bystander non-intervention. No mention of population frequencies of responders and nonresponders in the various bystander conditions was made in the introduction of their paper; instead, the psychological processes of individuals shifting responsibility, shifting blame, and assuming help from other bystanders were described. Darley

and Latané then radically trimmed and reshaped the model to fit into a number of simple independent–dependent variable relationships without realizing that this was the first step toward moving to a level of abstraction (i.e., frequencies) for which their model was not designed. Finally, following the accepted Pearsonian–Fisherian method of data analysis of the day, they addressed hypotheses regarding population parameters in which they had very little interest. The logical disconnect between their original model—based on individual psychological processes—and the statistical significance tests—based on ratios of simple frequencies obtained from numerous persons—was therefore complete. Subsequent studies in their program of research served only to solidify this divorce as they sought to replicate the abstract bystander effect without conducting any true replications. Their attempts to extend their understanding of bystander non-intervention were, in the end, haphazard because they had failed to develop an integrated model from which they could systematically descend increasingly deeper into nature. Ascending into the abstract realm of null hypothesis significance testing, Darley and Latané's incremental withdrawal from reality presaged the unfortunate fate of countless modern psychologists and all who have adopted this approach. It is perhaps a fate sealed with the lingering kiss of positivism, a relic of 19th-century mankind's attempt to elevate *scientia naturalis* to *scientia universalis*.

OBSERVATION ORIENTED MODELING

Observation oriented modeling offers an alternative to NHST and to the positivism of the Pearsonian–Fisherian tradition from which it sprang. As discussed previously and also in previous chapters, it is based on the philosophical realism of Aristotle and Thomas Aquinas that has found expression in the works of William Wallace (1996), Etienne Gilson (1990) and Jacques Maritain (1951). It is also largely consistent with the realism found in Rom Harré's thought and in the critical realism of the late 20th century (Harré, 1987).[12] Rather than focusing on the estimation of aggregate (e.g., means, standard deviations, and frequencies) population parameters that often have no empirical basis, attention is instead directed toward a world composed of things that have distinct, intelligible properties and powers and

[12] Also see Leplin's *Scientific Realism* (1984) for a collection of essays on critical realism. A number of contrary essays appear in this edited volume as well.

that interact with one another via time-ordered processes.[13] The starting point for the scientific study of bystander intervention must therefore be intelligible reality. As inherently intelligible, the scientist will develop a variety of concepts and causal mechanisms regarding nature, and these must be integrated in a model like the one shown in Figure 6.3. Simple variable-based models, like that shown in Figure 6.4, are of little value; in fact, the word "variable" is to be avoided as much as possible in observation oriented modeling. A variable is a concept—essentially a placeholder in the mind—and is not to be confused with the quality, attribute, event, etc. being studied. The terminology employed in observation oriented modeling helps to ensure that the focus is always on the things of nature and the explained ordering of observations found in an integrated model.

Despite its superiority, the model in Figure 6.3 is still rather anemic compared to the well-established models of the atom or cellular biochemical pathways, but this fact only serves to remind us that a great deal of painstaking work must be done to validate the model in its current form and then to fill in its specific details and extend its boundaries. Testing the model in its current form begins by recognizing how the observations are ordered. As can be seen in Figure 6.3, the model entails ordering the observations into the 0- and 4-bystander conditions and into the seeking help and failing to seek help conditions. The observations can be recorded simply as shown in Table 6.1. The data are not, however, to be conceptualized as independent and dependent variables but, rather, as two sets of ordered observations presented simultaneously in the table. The model clearly shows that the decision to seek help (the effect) is determined by the presence or absence of other bystanders (efficient cause). In the language of observation oriented modeling, the seeking help observations (ordered into "yes" or "no" units) should be "conformable" to the bystander condition observations (ordered into "0" or "4" units). The rotational analyses described in Chapters 2 and 3 can then be readily applied, yielding an overall percent correct classification (PCC = 76.92) equivalent to the observations that are consistent with the model; namely, 11 participants in the 0-bystander condition sought help and 9 participants in the 4-bystander condition failed to seek help

[13] Rom Harré (1970) summarizes such a viewpoint: "Science, conceived as the study of the behavior of things and materials and the elucidation of their natures, will be a dialogue in which the ascription of powers prompts the investigation of natures, and the analysis of natures leads to the discovery of new kinds of things and materials to which powers are again ascribed, restarting the cycle" (p. 314).

[(11 + 9)/26 = 76.92]. The individual classification strength index (CSI) values for the correctly classified observations are all above .90, and the *c* value is equal to .02 (1000 randomized trials), indicating that randomized versions of the 26 participants' ordered observations rarely yield a percentage as high or higher than 76.92. The multigram in Figure 6.5 shows the results in visual form, which generally support the model.

It must be understood that despite some similarities with the odds ratio and *z* test discussed previously, observation oriented modeling is distinct from these statistics because it does not involve the estimation of population parameters nor an appeal to a hypothetical normal distribution. The focus is instead on the two sets of ordered observations as entailed in the model in Figure 6.3 and an attempt to bring them into conformity via rotation of their deep structures. Randomized versions of the same observations are utilized to help judge the results of the rotation. The current *c* value is equal to .02, which represents a very small proportion, and it indicates the uniqueness of the conformable pattern of the two sets of ordered observations. Only in 20 out of 1000 randomized trials did the Observation Oriented Modeling software yield a frequency of correctly classified results as high or higher than the frequency obtained for the observations as they were recorded. Although it is common practice in NHST, an *a priori* cut-point or "significance value" such as .05 should not be used in observation oriented modeling. Instead, the *c* value should be considered in the context of the percentage of correctly classified observations, the CSI values, the misclassified observations, and the understanding that the consistency of the results must be demonstrated through exact replication. The *c* value simply

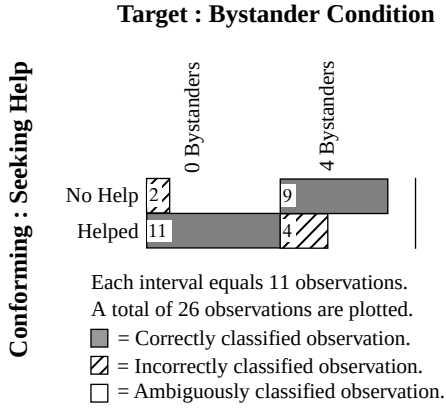

Figure 6.5 Multigram for the integrated model in Figure 6.3

provides an index of the unusualness of the obtained pattern of observations compared to randomized versions of the same observations. Recall from Chapter 4 that the *c* value is computed by randomly pairing ordered observations and then attempting to rotate them to conformity. In essence, the randomization procedure for the bystander study is working from an altered version of the model in Figure 6.3 in which the links connecting the number of bystander ordered observations to the seeking help ordered observations are completely arbitrary (random). One shuffled version of the ordered observations is therefore just as good as another and just as good as the actual pairings of observations. "Good" is here meant to indicate that the seeking help ordered observations can be brought into high conformity with the number of bystander ordered observations, yielding a high PCC result. Darley and Latané's results fared very well against these arbitrary pairings (*c* value = .02), but just how unusual a given PCC value must be is a matter of judgment that should be made on the basis of the model and the other numerical results and not simply on the basis of some arbitrary cut-point.

With observation oriented modeling, concerns regarding statistical assumptions and wide confidence intervals around population parameters with no empirical basis are replaced with concerns about individual observations. Both the model and the accompanying analysis are now on the same level—the level of the persons in this case. Any use of descriptive statistics, such as proportions for this study or means for quantitatively ordered observations, would serve a strictly summary function and would not be treated inferentially as in NHST. With all of this in mind, the current model classified 20 of 26 observations correctly, leaving 6 observations incorrectly classified. These latter 6 observations were 2 participants in the 0-bystander condition who did not immediately seek help and 4 participants in the 4-bystander condition who sought help in a timely manner. With observation oriented modeling, the goal is to improve the percentage of correctly classified results, which involves examining the 6 participants who were misclassified by the model. As noted previously, learning from anomalous or deviant observations is an important feature of scientific research and part of one's natural attitude when a realist philosophy of science is adopted. A convenient place to begin in observation oriented modeling is to examine the CSI values. In general, low CSI values for misclassified observations in the context of both a high PCC result and high CSI values for the correctly classified observations suggest that the few misclassified observations may represent a tolerable level of error. In the current case, however, the CSI values for the correctly and incorrectly

classified results are all high and equal to either .92 or .98. The uniformity of these CSI values is largely due to the simplicity of rotating observations that are ordered into only two units (0- and 4-bystander; seeking help and failing to seek help). More complex orderings of observations will be capable of yielding greater diversity in the CSI values.

The next step is to utilize other ordered observations that were made during the study in an attempt to clarify or expand the model in Figure 6.3. Darley and Latané did make many additional observations; for instance, they recorded the amount of time it took each student to seek help, they gathered demographic information, they interviewed the participants about their thoughts and emotions after the study, and they collected responses to personality questionnaires. Unfortunately, almost all of these observations were reported in summary statistics that cannot be utilized in observation oriented modeling, which is based on the original units of observation. The model cannot therefore be improved using the original observations made by Darley and Latané, which are of course unavailable after 50 years. Nonetheless, to demonstrate how model improvement would proceed, responses were fabricated for participants who stated they thought the seizure was fake. Recall that 20 of the 65 participants (31%) in Darley and Latané's study made this claim. Consistent with this proportion, 8 of the 26 participants (31%) in the 0- and 4-bystander conditions were considered to have regarded the seizure as fake, and these 8 individuals were split evenly between the two conditions. The fabricated observations are reported with the original data in Table 6.2.

With these new ordered observations in hand, the original model in Figure 6.3 must next be modified to include guessing. It is conceivable that a participant who guesses the seizure is fake will sit and wait in the room to see what else happens, or the participant may guess the seizure and seek help immediately because he or she may think the study is about morality or strength of character. Additional components would therefore be added to the model in order to adequately explain the behavior of the guessers. Two

Table 6.2 Contrived data for participants guessing the seizure was fake

		Sought help		Guessed seizure as fake	
		Yes	No	Yes	No
No. of bystanders	0	11	2	4	9
	4	4	9	4	9

simpler alternatives exist, however. To begin, the model in Figure 6.3 holds only for those participants who do not claim to guess the seizure is faked. This makes it clear that a conditional (logical) statement is involved in the attempts to improve the observation oriented modeling results: "If a given participant does not claim to guess the ruse, then the model in Figure 6.3 holds; else, an alternative model must be used." It is worth noting that this statement is made at the level of the ordered observations in the model—that is, at the level of the persons rather than aggregates. The first alternative is to follow this logic and simply exclude the guessers from the observation oriented modeling analysis. Doing so yields a PCC value of 94.44 (17 out of 18, c value $< .01$, 1000 randomized trials), CSI values all above .99, and the multigram shown in Figure 6.6. Only one participant, a student in the 4-bystander condition who sought help, was incorrectly classified.

A second option is to use logical statements in observation oriented modeling to create compound ordered observations from existing observations. Here, the logical statement is as follows: "If claim of guessing, then the unit of observation = 2; else, unit of observation = 0 for those who did not seek help or 1 for the those who did seek help." The assignment of "0," "1," and "2" is arbitrary and is not meant to represent a quantitative dimension or variable. In the language of observation oriented modeling, these values simply represent how the observations are ordered into three units (groups). These new ordered observations can then be rotated to conformity with the target bystander condition observations, yielding a PCC value equal to 80.77. Twenty-one of the 26 observations were

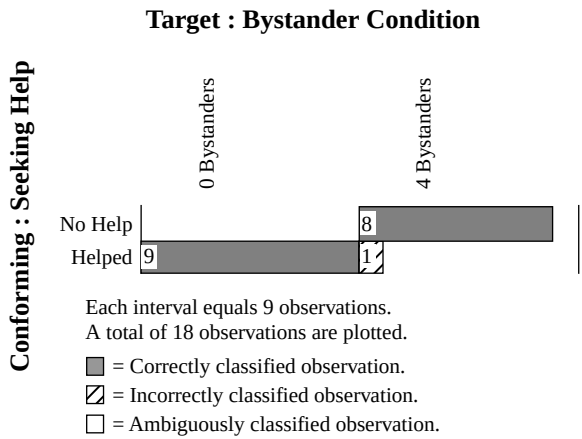

Figure 6.6 Multigram for participants who reported believing seizure

correctly classified with CSI values equal to .99 or 1. Four of the eight guessers were among the correctly classified observations, but the CSI values for all eight guessers were low and equal to .74. These CSI values represent the expected ambiguity associated with those who claimed to have guessed the seizure was fake. The *c* value was impressively low (.01, 1000 randomized trials), and the multigram in Figure 6.7 shows the results.

A third and perhaps best alternative is to work from the original model and recognize that Darley and Latané made the observations regarding guessing after the study was concluded—that is, after the participant had sought or failed to seek help in a timely manner and the experimenter had informed the participant of the study's design and purpose. Figure 6.8 shows a model in which the guessing ordered observations (yes or no) are logically connected with the seeking help ordered observations.

In observation oriented modeling, these two sets of ordered observations are crossed and then rotated to conformity with the 0- and 4-bystander condition observations. The results shown in the multigram in Figure 6.9 represent a numerical and conceptual improvement to the original model. The PCC value is 84.62 (22 of 26, *c* value $< .01$, 1000 randomized trials), and the CSI values are lowest for the guessers (.85–.88) compared to the nonguessers (.99–1.0). Again, these different CSI values represent the expected ambiguity (some may seek help, some may not) associated with those who claimed to have guessed the seizure was fake in the current model.

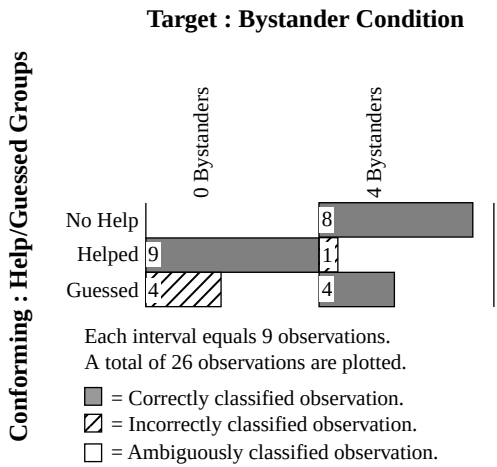

Figure 6.7 Multigram with "guessers" as a separate unit of observation

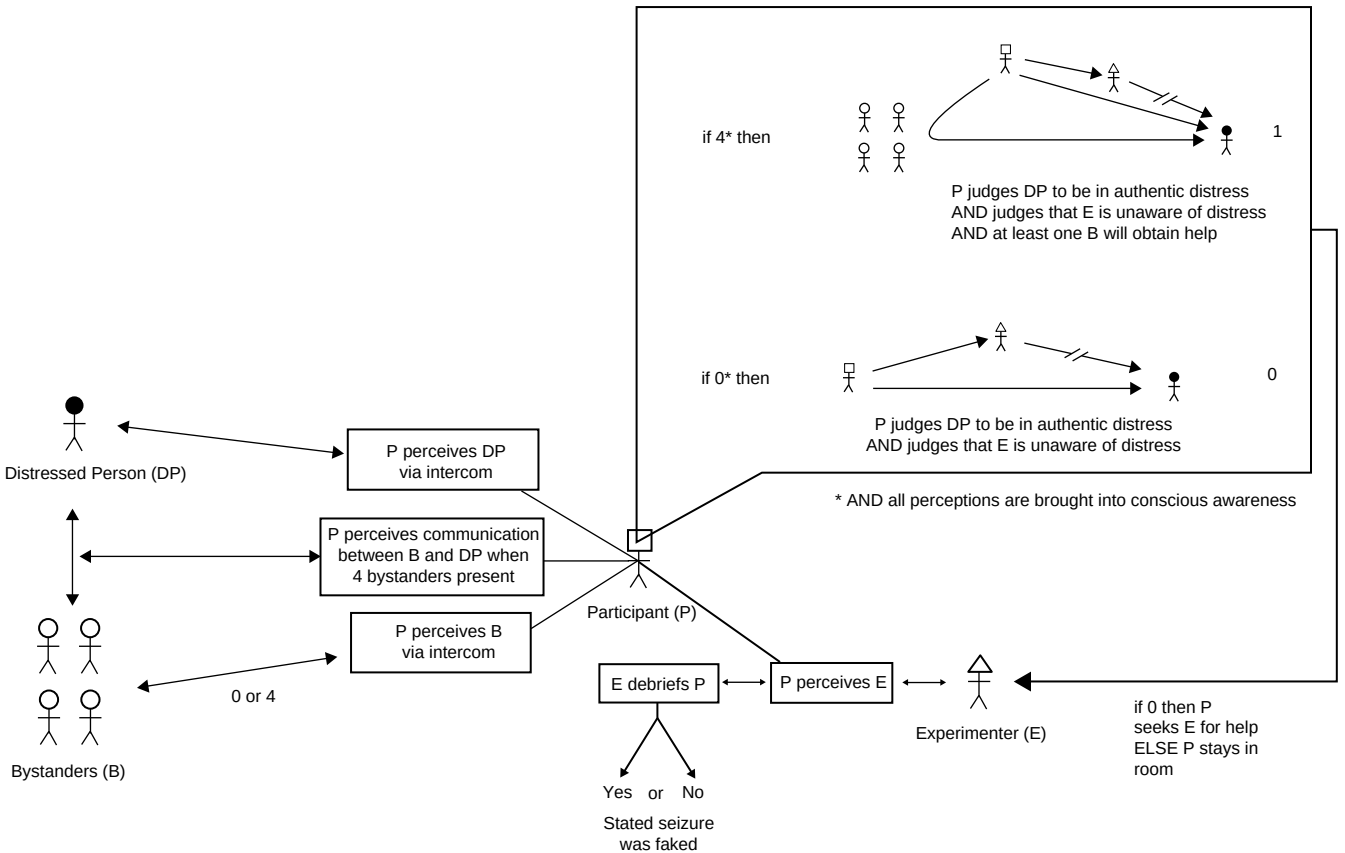

Figure 6.8 Revised integrated model of the 0- and 4-bystander conditions

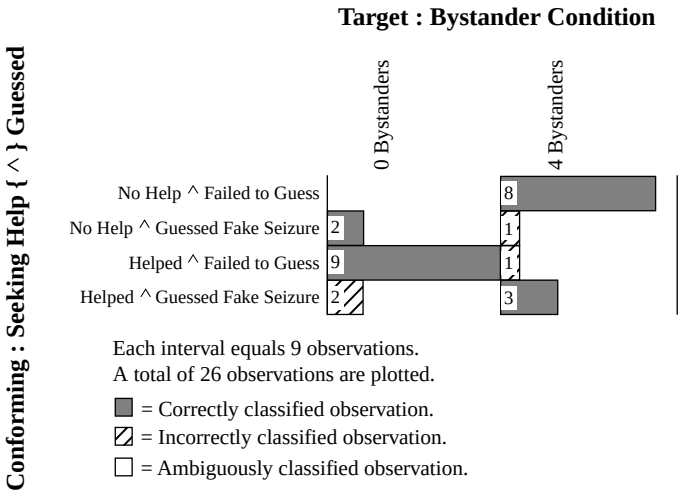

Figure 6.9 Multigram for logically connected guessing and helping ordered observations

Any one of the previously discussed approaches toward improving the model and observation oriented modeling results must of course be considered in the context of the philosophical ideas and technical matters presented. Most important, the *c* value—no matter how impressively small—is not to be considered as an index of replication. Results must therefore be subjected to an exact replication before further refinements to the model are made. An exact replication involves using precisely the same methods and procedures entailed in a given model to obtain an equal or greater number of independent ordered observations. If the results do not replicate, then the researcher must return to the drawing board. If the results do replicate, then the researcher can decide what steps to take to improve the PCC values by either refining the model's existing components and processes or expanding its range.

The psychologist using observation oriented modeling is thus reminded of the axiom, replication is a hallmark of *scientia naturalis*. Moreover, inherent in observation oriented modeling is the corollary that replication is valuable only in the context of modeling. A phenomenon may be perfectly reliable or repeatable and yet be completely misunderstood. It is through the use of a model that the psychologist ultimately demonstrates his or her understanding of the observations. These points were acknowledged by authors of several textbooks in the mid-20th century. Notably, Murray Sidman devoted 36 pages to the importance of exact (or direct) replications in his classic text

Tactics of Scientific Research (1960). Although his book focused mainly on the intensive study of individual organisms, detailing how experiments could be constructed and performed to obtain numerous observations from each, Sidman also discussed studies based on group data. He was careful not to confuse direct replication with systematic replication. The latter essentially involves using validated techniques and methods to extend one's model: "Where direct replication helps to establish generality of a phenomenon among the members of a species, systematic replication can accomplish this and, at the same time, extend its generality over a wide range of different situations" (p. 112). Systematic replication is moreover considered to be a gamble to the extent that one's methods, techniques, and models have not been thoroughly tested. In fact, "a psychology … whose investigators do not have adequate control over their primary datum, behavior, will be incapable of employing systematic replication in any extensive fashion" (p. 138). In the process of making a variety of questionable assumptions and estimating population parameters in NHST, it is not likely that a researcher will be focused on gaining the control of which Sidman speaks. Darley and Latané's work, published in the same decade as Sidman's book, clearly shows that some psychologists have long put the systematic replication cart before the direct replication horse. Darley and Latané established a statistical link between two variables in their original study. They then skipped any direct replication and instead sought to examine the so-called bystander effect in different situations. Without direct replication, they could not address the reliability of their methods and procedures or address clear threats to the validity of their conclusions, not the least of which is the fact that 20 (31%) participants claimed to have guessed the seizure as fake. With observation oriented modeling, attention is shifted away from the abstract realm of assumptions and probabilities in NHST to the observations as they are ordered by one's model. If the model is to be of scientific value, it must order observations on a consistent basis with a high degree of accuracy.

It is also instructive to examine Sidman's (1960) view of the relationship between the observed *p* value computed from NHST and the exact replication of a study:

> *The question of whether the experiment, if repeated, would yield a comparable difference between the experimental and control observations raises a second type of reliability problem—not the same problem as that involved in asking whether chance can account for the difference obtained in one experiment, but rather a problem in replication.*
>
> ***(p. 45)***

He clearly understood that NHST is a process of estimating population parameters that often have no empirical basis, and that these population parameters are aggregates (e.g., means, proportions, and variances). He rightly understood, then, that declaring a sample mean as statistically similar to a hypothesized population parameter is a form of inference that is distinct from the type of generalization psychologists hope to make. Most psychologists specifically wish to generalize their findings to other individuals not included in their studies. NHST, however, does not permit this type of generalization because the aggregate values on which it is based are derived from numerous individuals (excluding single-subject designs) and cannot in turn be used to make statements about any given individual. For example, the ordered observations for Darley and Latané's study revealed that 85% of the students in the 0-bystander condition sought immediate help for the seizure victim. The error made by far too many contemporary psychologists is to interpret this finding as indicating that the probability that Jonathan, Sue, Philip, or any other person will seek help in a similar situation is .85.[14] As another example, consider a psychologist who finds the mean for men on a mathematical reasoning test to be statistically higher ($p < .05$) than the mean for women. In the Discussion section of the published article, the psychologist is likely to make such statements as "Men outperform women on the test" or "Women seem to have difficulty processing the content of items on the reasoning test." These statements represent generalizations about particular classes of people (i.e., men and women), but the analysis represents an entirely different type of generalization—namely the generalization of a mean value computed from a subset of people (presumably selected randomly) to a population mean for the complete set of people. In order to be consistent with the analyses actually conducted as tests of the hypotheses, the psychologist should instead write, "The population mean for men on the test was estimated to be higher than the population mean for women." Statements such as this are not routinely found in published psychological studies because statistical analysis, particularly NHST, is thought to provide the justification for equating the two types of generalization.[15] For example, in an interview for a popular

[14] The logical difficulties involved in applying a frequency probability to an individual case are discussed in Chapter 9.

[15] Darley and Latané (1968) do little to quell these misunderstandings with statements such as "an individual subject is likely to respond if he thinks that others are present" (p. 380) or "male subjects reported the emergency no faster than did females" (p. 382).

newsletter, one noted psychologist claimed, "The goal of psychology is to make nomothetic laws—laws that apply to all people… The question is, how can you do that when you're sampling by convenience" (Feldman Barrett, 2005). The answer is that you cannot generalize to all people even when sampling at random and conducting NHST on data aggregated across persons. The statement conflates the application of universal laws to individual people with the estimation of population parameters through sampling techniques. In another instance taken from the *American Psychologist*, two noted researchers wrote, "However important universal principles are, they sometimes obscure how real people actually function. That is, by attempting to describe only the average, one runs the risk of describing nobody in particular" (Molder & Dweck, 2001, p. 192). These authors have unfortunately confused universals with means, which is tantamount to confusing the two types of generalization. The immediate cost of this confusion for these authors was that they proceeded to cite numerous aggregate- and NHST-based studies to support their argument regarding the importance of individual idiosyncrasies, never realizing that the methods employed in these studies contradicted their intention.

The more general and puzzlingly persistent errors in reasoning regarding NHST were documented in the previous chapter. With observation oriented modeling, psychologists may finally be able to gain their freedom from these subtle but costly confusions. It all begins with understanding the importance of one's integrated model, which serves as the explanation of the structures and time-dependent processes (i.e., the causes) of some natural system. Analysis is based directly on observations as they are ordered in the model, which yields two distinct advantages. First, when scientists make observations, they do so through particular apparatuses and measurement procedures following carefully scripted sets of operations. For instance, a chemist may weigh a compound using an electronic scale, being sure to follow the proper protocol for the scale, or a psychologist may score responses to a questionnaire using a scoring template provided by the test distributor. Observations are ordered under the constraints of these apparatuses and procedures, and observation oriented modeling recognizes their necessary and intimate relationship with modeling. Second, because the analysis is conducted on the observations as they are ordered by the model, it is conducted at the same level of abstraction as the model. Pragmatically speaking, this means that modern psychologists can escape the confusion that comes from analyzing aggregates while modeling individual persons or organisms. Observation oriented modeling provides a number of tools for

evaluating the adequacy of a model, including the PCC and CSI values and the *c* value. The overall goal is to correctly classify as many observations as possible with one's model and then to establish reliability of the model through exact replication. Generality of the model can then be established by extending its scope (i.e., adding more structures and processes) to account for increasingly more observations. Generalizations from a sample statistic to a population parameter are completely eschewed, and aggregate statistics are relegated to a secondary role—if they are utilized at all. Observation oriented modeling thus permits the psychologist to work in a more harmonious manner, focusing on the structures and time-dependent processes in his or her model using an analysis technique based on the patterns among ordered observations. In other words, the psychologist is directed toward the observable via his or her model and anchored in the observable via his or her methods.

CHAPTER 7

Modeling and Effect Sizes

Contents

Introduction 121
Aggregate Effects 122
Differences Between Means 123
Correlation and Shared Variance 136

INTRODUCTION

A task force composed of widely respected and well-known quantitative and experimental psychologists was assembled by the American Psychological Association (APA) in the 1990s. The task force was commissioned to offer a clear statement regarding the strengths and limitations of null hypothesis significance testing (NHST) and to provide guidelines for improving the quality of reported research methods and statistical analyses in psychological studies.[1] On the one hand, the APA's efforts were successful in bringing to the forefront a number of serious issues regarding NHST and the reporting practices of psychologists. Until the 1990s, criticisms of NHST were often published in books or journals with specialized audiences or limited readership. In contrast, the task force published its findings in the flagship journal of the APA, the *American Psychologist*, which is the most widely circulated social science periodical in the world (Wilkinson, 1999). On the other hand, the specific recommendations of the task force have had limited impact on the actual practices of research psychologists. Five years following the task force's report, Gigerenzer (2004) offered a sobering review of the continuing widespread and intransigent misunderstandings and misapplications of NHST.

One specific recommendation offered by the task force was the reporting of effect sizes to supplement the test of statistical significance (i.e., the p value). In the context of NHST, an effect size represents the magnitude

[1] Discomfort regarding NHST had been growing since at least the 1960s, but it was not until the 1990s that the APA would officially take notice. Jacob Cohen's 1994 paper, "The Earth Is Round ($p < .05$)," published in the *American Psychologist*, is considered as a catalyst for the formation of the task force.

Observation Oriented Modeling
ISBN 978-0-12-385194-9, Doi:10.1016/B978-0-12-385194-9.10007-6

© 2011 Elsevier Inc.
All rights reserved.

of an estimated population parameter (e.g., $\overline{x}$ as an estimate of μ), or the difference between two such estimates (e.g., $\overline{x}_1 - \overline{x}_2$ as an estimate of $\mu_1 - \mu_2$). The hope of the task force was that by focusing on effect sizes, psychologists would pay attention to the theoretical, practical, and clinical significance of their results in addition to statistical significance. The task force also hoped that a clear understanding of effect sizes would function as an antidote to the conceptual errors surrounding NHST and the *p* value. As noted, however, these hopes have not been realized because for all practical purposes the inclusion of effect sizes has only cosmetically altered the content of published psychological studies. One reason for this failure is that the task force remained wedded to the Pearsonian–Fisherian tradition and failed to offer any meaningful alternatives to study conceptualization and data analysis. As argued in Chapter 6, this tradition undermines the importance of exact replication and thwarts the development of integrated models that would support the accretion of scientific knowledge.[2] In this chapter, contrived data sets based on actual studies are used to demonstrate that the reporting of effect sizes fails to remedy these shortcomings. Observation oriented modeling is again shown to be an alternative to the Pearsonian–Fisherian approach that provides a more promising path to *scientia naturalis*.

AGGREGATE EFFECTS

The term aggregate has thus far been used in this book to refer to any statistic that summarizes a number of quantitative observations. In studies of psychology, the observations are often made on different people, although an aggregate can certainly be computed from multiple observations of a single person. The only requirement is that the observations are obtained from a common scale (e.g., inches, centimeters, or millimeters) or represent countable objects, behaviors, etc. in a common class (e.g., males, females, eye blinks, or key presses). By far the most popular aggregate in psychology

[2] It is unfortunate that the APA task force failed to address the importance of replication in its recommendations. Its report contains entire sections on statistical assumptions, statistical power, random assignment, measurement, and so on, but there is no section devoted to replication. It is particularly unfortunate because Cohen closed his famous 1994 *American Psychologist* article with the following: "As I noted, we have a body of statistical techniques that, used intelligently, can facilitate our efforts. But given the problems of statistical induction, we must finally rely, as have the older sciences, on replication" (p. 1002).

is the arithmetic mean, followed closely by the variance and standard deviation. The arithmetic mean is simply the average of a set of numbers, whereas the variance and standard deviation quantify the extent to which the numbers are different—that is, the extent to which they vary in magnitude. Effect sizes in psychological studies are therefore typically expressed in terms of differences between means or in terms of the percentage of overlapping variance between two sets of quantified observations.[3] Moreover, effect sizes are almost always computed in the context of NHST and considered to represent population parameters rather than sample-specific quantities, a point clearly supported by the claims of meta-analysts. It was shown in Chapter 6 that the process of estimating population parameters can actually thwart the development of sound scientific models. This problem is magnified in psychology because its models are rarely constructed to explicitly predict means, standard deviations, proportions, or any other type of aggregate statistic; however, aggregate statistics are at the center of most analyses intended to support or refute models in psychology, and NHST is regarded as the "backbone of psychological research" (Gerrig and Zimbardo [2002, p. 46] as cited in Gigerenzer [2004, p. 589]). Well-intentioned models of human action are therefore unconsciously traded for ill-conceived variable-based models, and following the APA's advice by reporting effect sizes does little to nullify the consequences of this perfidious swap.

DIFFERENCES BETWEEN MEANS

In response to the increased emphasis on effect sizes that culminated in the APA's task force, psychologists have sought to justify the typically small effects (e.g., small differences between means) reported in their studies. Deborah Prentice and Dale Miller (1992), for instance, argued that a statistical index of effect size is not sufficient for demonstrating the importance of an experiment's outcome. Other criteria must be used as well, such as the theoretical importance of the effect or its ostensible pervasiveness across different research designs. They cited and discussed a variety of studies to support their claim, including a series of studies published in 1980 that investigated the hypothesized tendency of people to favor members of their own group over members of other groups (Locksley,

[3] An excellent summary of various aggregate indices of effect size is provided by Rosnow and Rosenthal (2003). Their references provide numerous relevant citations on this topic.

Ortiz, & Hepburn, 1980). In the first of these studies, 18 university students, tested in groups of 6, privately drew tickets from a can that would allegedly assign them randomly to one of two groups, Phi or Gamma. The participants also understood there would be 3 students in each group; however, unbeknownst to the students, all were assigned to the Phi group. After their assignment, the 6 students sat in private cubicles and were asked to imagine allotting up to 100 coins to each of the other 5 members, 3 of whom were Gammas (the "out-group") and 2 of whom were fellow Phis (the "in-group"). The allotment for each Phi or Gamma was written on a separate piece of paper, and each student was not permitted to allot coins to himself or herself. The researchers then averaged the number of coins allotted to the in- and out-groups and found the mean was higher for the former ($M = 85.3$) than the latter ($M = 64.9$). Following the standard operating procedure for the Pearsonian–Fisherian approach, the difference was tested with a dependent samples t test and judged to be statistically significant: $t(17) = 2.91$, $p < .01$.

Locksley et al. (1980) interpreted this finding as supportive of a fundamental bias that leads people to favor others in their group, even if the group is formed by completely arbitrary means (i.e., by randomly selecting tickets). Prentice and Miller (1992) went even further and referred to the result and others like it as "impressive demonstrations of ethnocentrism" (p. 161). Such a claim seems sensational because it is not clear how the study had anything to do with well-known ethnic biases, but it does suggest that the causal mechanism underlying the allotment of the coins to fellow Phis is the same mechanism underlying ethnocentric behaviors. What exactly are the causal mechanisms underlying the coin allotment study? Unfortunately, no integrated model was proposed or tested, and in fact the authors of the study suggested at least three potential causal explanations based on prior, similar studies: assumed similarity of personal characteristics for in-group members, reward expectation from in-group members, and out-group discrimination (Locksley et al., 1980, pp. 774–775).[4] Figure 7.1 shows how a generic in-group biasing mechanism might be represented in an integrated model for the study. The

[4] Locksley et al. (1980) state, "Though the minimal intergroup discrimination effect has proven highly reliable, research attempting to determine the basis for the effect remains inconclusive" (p. 774). In other words, the causal mechanism has not been determined. As has been argued in this book, in the context of NHST and the Pearsonian–Fisherian tradition, this admission is not surprising.

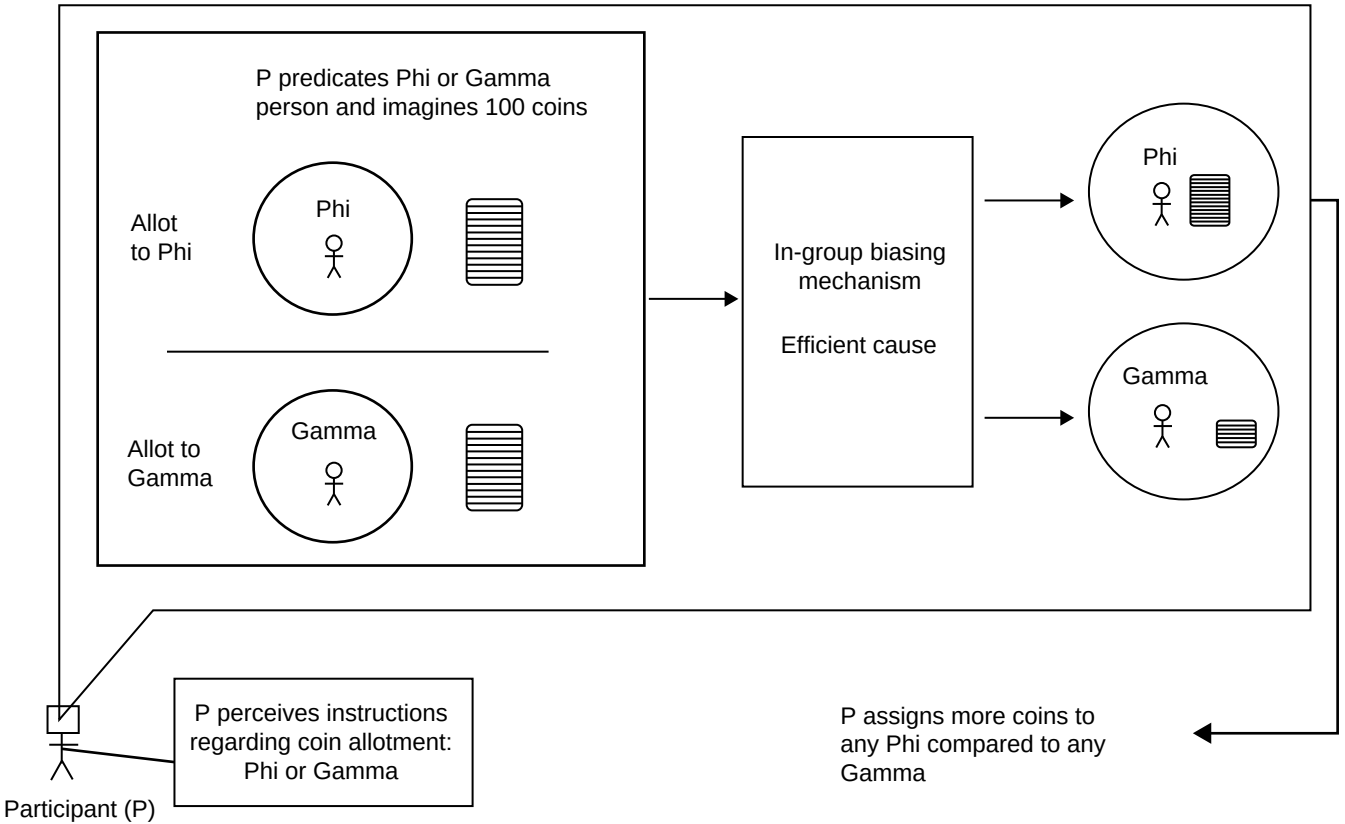

Figure 7.1 Efficient cause integrated model for coin assignment portion of in-group/out-group study

participant (P) is shown to be allotting imagined coins to either a Phi or a Gamma, separately, on each sheet of paper. After predicating the target person as a Phi or a Gamma and imagining the coins, the in-group biasing mechanism is shown to operate as an efficient cause that generates further judgments resulting in more coins for each Phi than for each Gamma. By asking the participant to place each allotment on a separate sheet, the study authors are implying that each judgment is independent yet driven by the same efficient cause mechanism. Figure 7.2, however, shows a different causal process not foreseen by the researchers. As can be seen, the student participant is seated alongside five other participants, two of whom are considered to be fellow Phis and three of whom are considered to be Gammas. The exact instructions read by the student simply stated that five sets of 100 coins were to be allocated to the other five persons, and that each person did not have to receive the same number of coins. The figure shows that the participant is aware of the other students in the room. The participant is not certain exactly who the three Gammas are, but he or she is certain there are three Gamma persons and two fellow Phi persons. If the participant construes (or predicates) group equality, then he or she will wish to assign the same number of coins overall to both groups. This construed ideal state represents a final cause to which the assignment of coins will be brought into conformity. Specifically, the two fellow Phis will receive more coins because the participant (who is a Phi) will not receive any coins. For example, the student might assign 90 coins to each of the two Phis and 60 coins to each of the three Gammas, thus giving an equal number of coins (180) to each group. It is perhaps telling that the ratio of 90/60 is close to the ratio of means, 85.3/64.9, reported in the study.

The final cause model in Figure 7.2 is a more holistic representation of the experiment because it recognizes the participant's awareness of the other people and their assignment to groups as well as the participant's inherent nature as a person to reach a desired end. The competing interpretation of the reported results compromises the conclusion of the experiment, or at least renders the results as ambiguous. Prentice and Miller (1992) made no such criticisms of the study because their argument revolved around effect sizes and significance testing within the Pearsonian–Fisherian tradition. Locksley et al. (1980) were similarly locked into a variable-based modeling approach that prevented them from sketching the causal mechanisms that might underlie their own study. Such a sketch may have helped them to realize that, at a minimum, the study needed to be redesigned so the allotment of coins would be based on equal numbers of persons in each group.

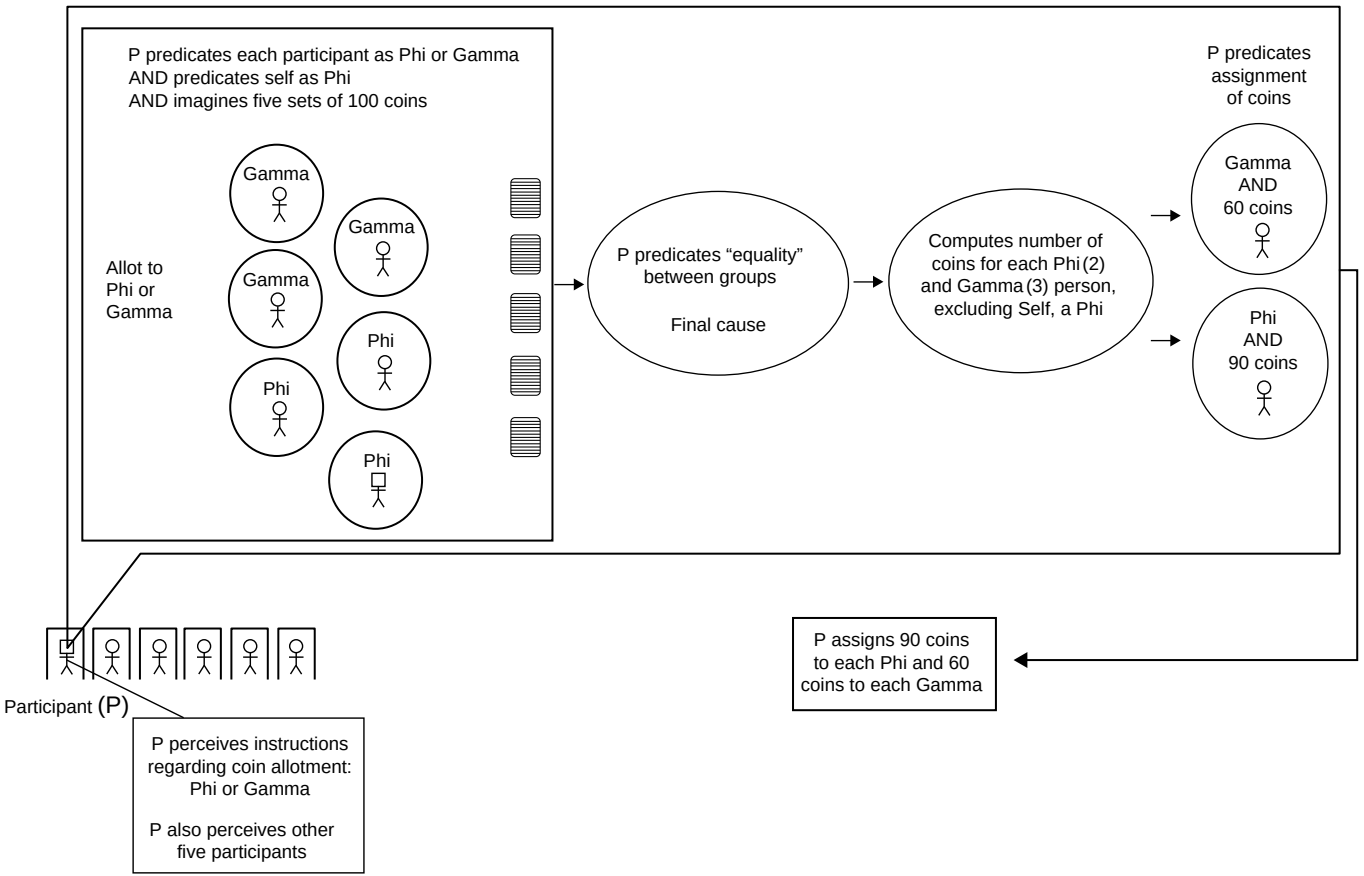

Figure 7.2 Final cause integrated model for coin assignment portion of in-group/out-group study

It is also instructive to examine the second experiment conducted by Locksley et al. (1980) and published in the same article. They set up the experiment by stating, "Experiment 2 replicated Experiment 1," but then in the same sentence added, "using a different dependent variable" (p. 776). As discussed at length in Chapter 6, such language fails to recognize the difference between exact replication and extensive replication. Taken by itself, the word "replication" infers an exact replication, which is not what the researchers undertook in their second experiment. In the realm of variable-based modeling, however, the two types of replication are fused because what is being replicated is a relationship (presumed to exist in an abstract population) between two equally abstract variables. This is the so-called "effect" being replicated in Experiment 2. Such reasoning is surely confusing and gives the impression that psychologists are utilizing the language of science in an esoteric and possibly corrupt way.

The term "effect size" connotes a causal explanation ("cause and effect"), and careful examination of exactly how effect sizes are computed shows that such a connotation can prove dangerous. Consider Experiment 2 referred to previously. In this study, 17 student participants were again asked to draw tickets from a can that allegedly assigned them to the Phi or Gamma groups.[5] As in the first study, all of the participants in fact drew Phi tickets but were led to believe that the number of Phis and Gammas equaled 3 persons per group. Instead of allocating coins, however, the participants now rated the 3 other Gamma and 2 Phi participants on five desirable and five undesirable traits (e.g., good-natured and self-centered) using a 5-point Likert-type scale ranging from 1, "never or almost never true," to 5, "always or almost always true." Each person's Phi and Gamma undesirable trait ratings were reflected and then summed separately with the desirable trait ratings. Two sum scores ranging in value from 10 to 50 were thus obtained for each of the 17 participants.[6]

Using a dependent samples *t* test, Locksley et al. (1980) tested the null hypothesis that the population mean difference between the alleged Phi and Gamma persons was exactly equal to zero (H_0: $\mu_{Phi\ -\ Gamma} = 0$). The

[5] The participants were tested in groups of six individuals, but one participant failed to show up for the experiment. The authors note that a research confederate filled in for the missing participant. Obviously, ratings were not collected from the confederate.

[6] As in the first study, differing numbers of Gammas and Phis may have biased the results. Averaging three Gamma ratings compared to two Phi ratings may have at least led to greater variability within the latter sum scores.

alternative hypothesis was that the population mean difference was not equal to zero (H_A: $\mu_{Phi\ -\ Gamma} > 0$ or $\mu_{Phi\ -\ Gamma} < 0$), and the authors specifically expected the difference to be positive. The results supported the variable-based model and the researchers' hypothesis because the mean for the Phi group ($M = 36.6$) was greater than the mean for the Gamma group ($M = 33.5$), and the result was statistically significant: $t(16) = 2.22$, $p < .05$, two-tailed.

For pedagogical purposes, a data set that mimicked these results was created with $M_{Phi} = 36.6$, $M_{Gamma} = 33.5$, $t(16) = 2.24$, $p < .05$, two-tailed. The box plot shown in Figure 7.3 reveals no outliers, although it does show a substantial amount of skew, which is ignored due to the presumed robust nature of parametric tests to non-normality.[7] Effect sizes were rarely reported in the 1980s, but had Locksley et al. (1980) done so, they would likely have used Cohen's *d*, which is the standardized difference between two means.[8] Each participant contributed individual ratings for the Phi and Gamma persons; consequently, the observations were not independent. This lack of independence is taken into account in the analysis through the use of difference scores computed from the two sets of ratings. The mean ($\overline{x}_{diff}$) and standard deviation (s_{diff}) for these difference scores for the current data were 3.12 and 5.73, respectively. It is upon these aggregate statistics that Cohen's *d* is based:

$$d = \frac{\overline{x}_{diff}}{s_{diff}} = \frac{3.12}{5.73} = .54 \qquad [7.1]$$

Clearly, Cohen's *d* is equivalent to a *z* score based on aggregates. In general, it can range in value from $-\infty$ to $+\infty$, with $d = 0$ indicating a mean difference of 0, or equivalently, no difference between the two

[7] The robust nature of parametric tests, even to violations of the normality assumption, is a matter of debate. For instance, in *Statistical Methods for Psychology*, David Howell (2007, pp. 203–204) outlines the generally robust nature of *t* tests to violations of homogeneity and normality but also reviews criticisms of this belief. It is also not difficult to find statements such as the following in the published literature: "Within the context of analysis of variance (ANOVA), it has been well established that heteroscedasticity, skewness, and outliers (i.e., a point, or points, far from the central point in the distribution) can have a devastating effect on both Type I errors and power" (Keselman, Algina, Lix, Wilcox, & Deering, 2008, p. 110). There can be no debate, however, that the normality assumption is ignored in the vast majority of published psychological studies.

[8] The central citation for effect sizes and statistical power is Jacob Cohen's *Statistical Power Analysis for the Behavioral Sciences* (1969/1988).

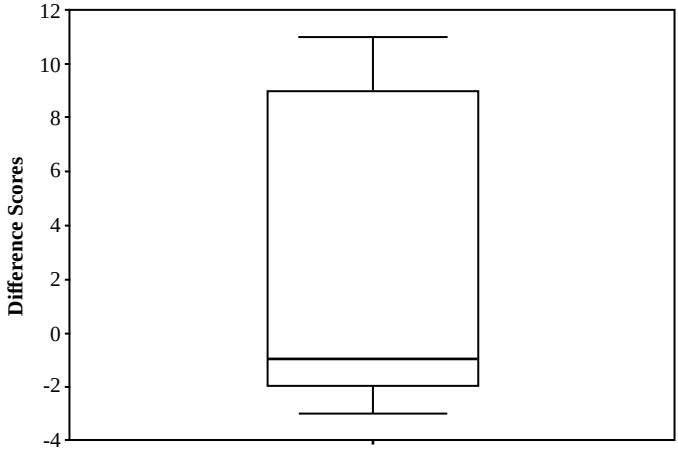

Figure 7.3 Box plot for difference scores computed from Phi and Gamma trait ratings

means under consideration since $\overline{x}_1 - \overline{x}_2 = \overline{x}_{diff}$ (but $s_1 - s_2 \neq s_{diff}$). Cohen developed rules of thumb or cut-points for absolute values of *d*: .2 = small effect, .5 = medium effect, and .8 = large effect. His justification for these values was based on the magnitudes of effects he catalogued in his review of published psychological studies as well as on the proportion of overlap between two normal distributions of the mean; for example, for $d = .8$, two distributions will overlap 53%, whereas for $d = .2$, they will overlap 85% and show weaker support for the alternative hypothesis. When *d* is equal to 0, the two distributions will overlap completely, corresponding to the two population means being equal (e.g., $\mu_{Phi - Gamma} = 0$). These cut-points and proportions naturally depend on the usual assumptions of parametric tests, such as equal population variances and normal distributions.

Cohen's *d* has become popular in recent years for at least two reasons. First, because it is a standardized statistic, it can be used to compare aggregate effects across studies. Consider another researcher who follows this study using similar procedures for assessing trait judgments but who employs a different rating scale to collect the ratings. Suppose this researcher uses a scale that ranges from 1 to 10, and that the mean difference between the summed Phi and Gamma ratings is equal to 7.0 ($s_{diff} = 12.0$). Does this larger difference represent a truly larger average effect resulting from the experimental procedures, or is it simply a consequence of using a scale with more options? The result yields a *d* value equal to .58, which is only slightly

higher than the result in Eq. [7.1]. The two average effects would therefore be considered as highly similar, and in a meta-analysis the values would be averaged to obtain an *ostensibly* more accurate estimate of the standardized difference between the two population means. "Ostensibly" is emphasized here because in addition to the well-known file drawer problem, mean differences established via meta-analyses are not always replicable, and differences between studies that are not usually considered important may in fact impact the estimated effect sizes severely.[9]

The second reason for the popularity of Cohen's d is that it provides researchers with a convenient heuristic for judging their statistical results as "small," "medium," or "large" in the context of NHST. Much like using "$p \leq .05$" as a convention for making a binary decision regarding the presence or absence of an effect, Cohen's cut-points have become conventions for judging the magnitude of an effect. On the one hand, computing d and interpreting according to Cohen's conventions satisfies the recommendations of the APA task force for reporting estimated effect sizes. On the other hand, by all appearances, it seems that one "ritual" ("$p \leq .05$ indicates that my results are important and publishable") has been replaced with—or at best supplemented by—another ("$d \geq .8$ indicates that my effect is large and theoretically, clinically, or practically important"). Evidence for this assertion can be found in innumerable studies published in approximately the past 5 years that report and interpret d using Cohen's

[9] The so-called file drawer problem refers to unpublished studies that cannot be included in a meta-analysis because of their inaccessibility. When collating the results to determine the statistical association between, for instance, self-esteem and grade point average, how can the meta-analyst gain access to the unknown number of unpublished studies on this topic residing in researchers' file drawers? Advocates of meta-analyses have created various statistical adjustments for determining the alleged impact of unpublished studies on the estimated effect sizes (Rosenthal & Rosnow, 2008). The legitimacy of these adjustments, however, has been seriously challenged (Scargle, 2000; Schönemann & Scargle, 2008). Advocates of meta-analyses must also confront the fact that a review of meta-analyses in the medical literature showed that double-blind experimental studies failed to replicate prior estimated effects for 35% of the results published in meta-analytic studies that were later tested (LeLorier, Genevieve, Benhaddad, Lapierre, & Derderian, 1997). Finally, in a white paper by Paul Barrett, the statistical associations between different personality traits were found to vary across questionnaires used to assess the traits. This finding contradicts the assumption often made in meta-analyses that differences in measurement devices are superfluous as long as they assess the same "variables" (Barrett & Rolland, 2009). In most general terms, meta-analyses are conducted to estimate population parameters and are thus anathema to observation oriented modeling.

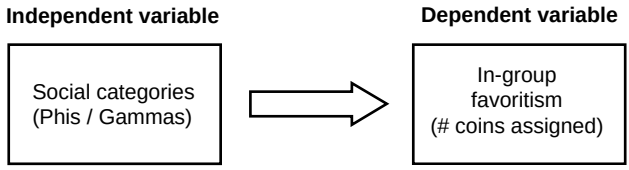

Figure 7.4 Variable-based model for in-group/out-group study

conventions and then fail to provide any meaningful discussion of the effects in the study conclusions. To the extent that such discussion can be found, it typically takes on the form of simply reiterating the .2 = small, .5 = medium, and .8 = large interpretation of the estimated effect size.[10]

The primary reason this new ritual has emerged is because psychologists routinely work with simple variable-based models when analyzing their data. Such models do not generally lay out the structures and time-dependent processes (i.e., the causes and effects) intrinsic in any psychological theory. For the first study discussed previously, the model is shown in Figure 7.4. Consistent with this variable-based model, a *t* test was used to determine whether or not the link between the two variables was consistent with the null hypothesis (H_0: $\mu_{Phi\ -\ Gamma} = 0$), and the primary focus of the analysis was therefore on estimating the magnitude of the difference between population mean ratings. However, the causal mechanisms underlying the actual Phi and Gamma ratings do not reside at the level of population means. Instead, these mechanisms—like those shown in the integrated models in Figures 7.1 and 7.2—exist only at the level of the persons. Any discussion of effect size must therefore be made at the level of the persons in this example study, not at the level of means computed across persons.

Contrary to the aggregate approach, observation oriented modeling works at the level of the observations. Analysis of the previously presented data from this perspective would ideally begin with an integrated model that expresses the hypothesized causes and effects at the level of each student participant. Such a model would indicate that each participant should rate each Phi student as more desirable than each Gamma student. In their current form, however, the observations are aggregated for each student rater. Nonetheless, the average ratings will be analyzed "as is" and compared

[10] See Thompson (2004). A number of informative papers regarding the pitfalls of significance testing and the obstinate behaviors of researchers are reported in Volume 33 of the *Journal of Socio-Economics* (2004).

to the previous results in order to clearly demonstrate the differences between NHST and observation oriented modeling. Following the hypothesis, then, the difference between the average Phi and Gamma ratings should be positive for each student, thus indicating a more favorable view of the former. The observed mean differences for the 17 participants show a wide variety of judgments: −3, −3, −3, −2, −2, −1, −1, −1, −1, 4, 7, 8, 9, 10, 10, 11, and 11.

In all of the examples from previous chapters, both target and conforming observations were available and could be rotated using the methods described in Chapter 3. Here, only one set of observations is available, and an alternative route must be traveled. Specifically, frequencies of the actual observations are compared to expected frequencies of the same units of observation. Considering the study methodology and sum scores ranging in value from 10 to 50, a total of 81 units of observation are possible (i.e., scale point differences ranging from −40 to 40, including 0). The study clearly lacks a sufficient number of observations to even model this number of units, and the lack of an integrated model leaves us with little guidance. Consequently, the small number of observations were assigned to three units that matched the hypotheses: Phi > Gamma, Phi = Gamma, and Phi < Gamma. As stated previously, the expectation is that the observations for each participant will match the "Phi > Gamma" pattern. No observations were expected for the "Phi < Gamma" and "Phi = Gamma" units.[11] Comparing the trichotomous actual and expected observations in the Observation Oriented Modeling software yields the multiunit frequency histogram in Figure 7.5, which shows that only 8 of the 17 observations matched expectation. The "Phi < Gamma" unit accounted for the other 9 observations, contradicting the researchers' hypothesis at the level of the persons in the study despite the opposite, medium-sized effect indicated by Cohen's *d*. Because the observations can be expressed in deep structure form, a randomization test can also be performed that compares the

[11] Blocking and dichotomizing are widely regarded as poor practices among social science researchers. Gavan Fitzsimons offers the most pointed criticism of dichotomizing in his editorial, "Death to Dichotomizing" (2008). However, the arguments against such practices are centered on issues of statistical power, assumed continuous quantitative structure, and the estimation of population parameters, none of which apply to observation oriented modeling. Ideally, an integrated model will guide the observations, but practical issues, such as insufficient numbers of observations in the current example, may necessitate a pragmatic regrouping of units.

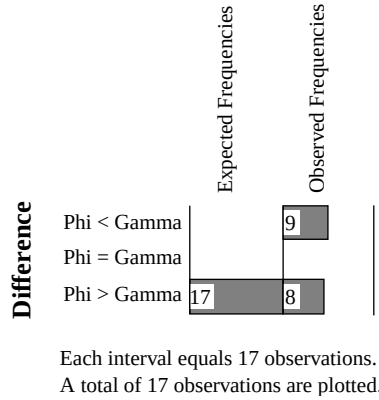

Figure 7.5 Multigram for signed difference between Phi and Gamma persons

expected and actual unit frequencies with regard to the number of overall matches. The resulting *c* value for the current data was surprisingly low, .17, suggesting a relatively unusual match between expected and observed frequencies. As can be seen in Figure 7.5, however, the *c* value is in a way benefiting from the empty "Phi = Gamma" unit, which also has an expected frequency of zero. Deleting this unit from the analysis yields a *c* value of .68, indicating chance agreement between expected and observed frequencies.

Observation oriented modeling thus provides a more accurate picture of the results, and this example shows that Cohen's *d* is an aggregate statistic of "effect size" that represents the standardized difference between two means. It does not indicate the typical, or most frequently occurring, difference between pairs of observations in a study. The psychologists conducting the previously discussed experiment deceived students into believing they were separated into two groups in order to generate an effect in the participants' subsequent ratings of one another. The cause-and-effect sequence—primarily understood as efficient cause—resides at the level of the persons in the study, not at the level of the aggregate (i.e., the arithmetic mean). The "effect" in the effect size estimate, $d = .58$, must therefore not be confused with the effect of each person rating his or her fellow Phi students in a more favorable manner than the Gamma students. Again, Cohen's *d* is a standardized difference between means; it is not an indication of the "typical effect" resulting from the experimental manipulation in a study. Locksley et al. (1980, p. 777) failed to see this distinction when they stated, "An analysis of the

trait ratings revealed that subjects were significantly more likely to rate Phis in a socially desirable direction than they were Gammas" (note the allusion to subjects), even though the effect they tested with NHST was based on aggregates (see Eq. 7.1). The current data starkly demonstrate this distinction by revealing a medium effect, according to Cohen's conventions, consistent with the researchers' expectations while simultaneously revealing that, contrary to expectation, a slight majority of participants (9 of 17) rated the Gamma students more favorably than the Phi students. It was only by paying attention to the observations that this contradiction was made obvious, resulting in the correct conclusion of no systematic effect; in other words, the participants did not show a bias in favor of in-group members.

Perhaps the data from the original study were not actually structured in this way. The current data were, after all, fabricated to demonstrate a reversal in conclusions between NHST and observation oriented modeling. Defenders of the Pearsonian–Fisherian tradition, such as Prentice and Miller (1992), might therefore attempt to find solace in speculating that such data patterns rarely occur in practice, particularly with large differences between means. There can be no doubt that when such differences are large and the standard deviations of the variables are small, the generation of such contradictory data becomes impossible. As noted by James Hemphill (2003), however, it appears that approximately two-thirds of published studies in psychology report effect sizes (as correlations) that are medium or small according to Cohen's original cut-points. This fact, coupled with the confusion regarding replications, completely undercuts Prentice and Miller's attempt to circumvent the effect size issue. There remains an inherent ambiguity when attempting to make general statements about individual observations on the basis of aggregate statistics, and this ambiguity is exacerbated as the effect sizes grow smaller. It is therefore difficult to say what has generally been learned about the human psyche regarding in-group bias or sundry other actions and qualities from aggregate, variable-based models.

With observation oriented modeling, this particular obfuscation is completely avoided because the psychologist works in greater harmony with the observations as they are ordered into units by his or her model. With an integrated model in hand, even in simple form, the researcher also never loses sight of the units of observation employed and is ready to focus not on aggregate statistics but instead on patterns that explain individual observations. For the specific study of in-group bias, the starting point

would thus be an integrated model positing more details about the material, efficient, formal, and final causes underlying coin allotment and trait ratings.

CORRELATION AND SHARED VARIANCE

Another class of effect sizes utilized frequently by psychologists includes both correlations and squared correlations. Imagine a study, for instance, in which a "judge" participant is asked to rate the degree of narcissism in 50 separate photographs of fellow students. Following the Pearsonian–Fisherian tradition and contemporary personality research methodology, the judge is asked to rate each photograph using a 7-point scale ranging from 1, "not at all narcissistic," to 7, "extremely narcissistic." The students in the photographs also provide self-ratings of narcissism using the same scale. The goal of the research is to address the accuracy of the raters' judgments with the belief that narcissism can in fact be reasonably identified from only a photograph. The ratings are typically presumed to represent an underlying continuously structured quantity, narcissism (but see Chapter 8), and Pearson's correlation coefficient, r, is therefore typically used with such ratings. Ratings were contrived for this example with a statistically significant correlation equal to .30, $p < .05$, two-tailed.

The properties of Pearson's r must be understood in the context of NHST. An appropriate starting point is Figure 7.6, which shows that the correlation coefficient, like most statistical techniques utilized by psychologists, pertains to a variable-based model. For this study, the variable "self-reported narcissism" is linked to the "judged narcissism" variable in a presumably linear manner. The link is also represented as an arrow to indicate that the goal is to predict the judge's ratings on the basis of a linear function of the self-ratings. The connecting link in Figure 7.6 is quantified as $r = .30$, which is considered to be an estimate of the population correlation coefficient, ρ. The null hypothesis in NHST is stated as $\rho = 0$, and the alternative is stated as $\rho > 0$ or $\rho < 0$.[12]

[12] The assumptions in NHST underlying Pearson's correlation are as follows: bivariate population distribution, independence of pairs of observations, random sampling, linear association between variables, continuous variables, and homoscedasticity. Arguably, the independence assumption is violated for studies such as the one described here because the 50 ratings were all made by the same judge. It is difficult to reason exactly how this particular violation would bias the statistical test and increase the Type I error rate, which perhaps explains why it is routinely ignored.

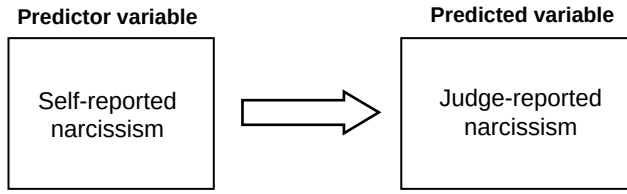

Figure 7.6 Variable-based model for predicting narcissism

Pearson's correlation coefficient is the average of the cross-products of the z scores for the two variables, labeled x and y, and is therefore an aggregate statistic:

$$r_{xy} = \frac{\sum z_x z_y}{n-1}, \qquad [7.2]$$

where n represents the sample size. It is inherently standardized with a set range—namely -1 to $+1$, with zero indicating no linear relationship. Because of this known range, the correlation coefficient, at first glance, serves as an efficient index of effect size. For instance, a correlation of .95 would clearly indicate a more impressive linear relationship between two variables than would a correlation of .15.

Another way to conceptualize Pearson's r is as a standardized covariance. The covariance statistic is a measure of linear association that can range in value from $-\infty$ to $+\infty$, with zero again indicating no linear relationship between the two variables under consideration. Because it does not have a set range, the covariance is not as readily interpretable as the correlation. For example, a covariance equal to 517 for the model in Figure 7.6 would indicate a positive linear association between the two variables, and it may seem quite impressive; however, in standardized form, it may convert to a small correlation equal to .15. It is useful to think of the correlation coefficient as a standardized covariance because it can be verbally described as indicating the extent to which two variables co-vary. In other words, the correlation can be thought of as an index of co-variability between two vectors of standardized numbers and is hence properly considered as a variance-based statistic. This fact is readily apparent when Pearson's r is squared, yielding a statistic that indicates the proportion of overlapping variance between two variables. The square of the previous correlation (.30) equals .09, meaning the judge and self-reported narcissism ratings share 9% of their variance. This proportion of overlap can be represented in a Venn diagram like the one shown in Figure 7.7.

The squared correlation, historically referred to as the coefficient of determination, is an aggregate statistic, and it is probably the most popular

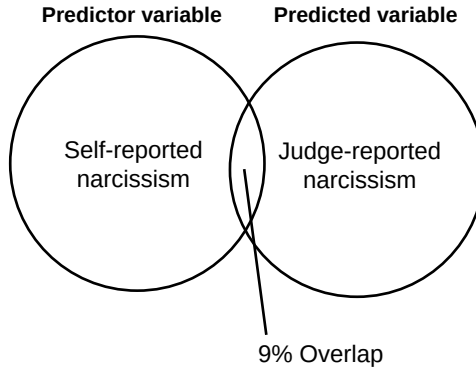

Figure 7.7 Venn diagram showing r^2 as proportion of overlapping variance

measure of effect size available to psychologists, who find it easier to comprehend the meaning of a proportion rather than a correlation or standardized difference between means (e.g., Cohen's *d*). Part of its appeal may also stem from the asymmetrical nature of most psychological variables. The arrow in Figure 7.6 shows that the goal is to predict the judge's narcissism ratings from the self-reported ratings, or to explain the former in terms of the latter. Consequently, the squared correlation might be described as indicating that the self-reported levels of narcissism explained 9% of the variance in the judge's ratings, which seems to be a small amount, leaving 91% of the variance unexplained.

Despite its intuitive appeal, however, a number of methodologists have argued that r^2 should not be preferred as an index of effect size. These critics instead support reporting the original correlation coefficient and interpreting it normatively or relying on entirely different metrics to convey the importance of a statistical effect. With regard to the first strategy, Cohen popularized the establishment of conventional cut-points for interpreting effect sizes. These cut-points are usually established through reviews of the relevant research literature. Hemphill, for instance, devised the following criteria based on meta-analytic reviews of the literature: $|r| < .20$, small effect; $.20 \leq |r| \leq .30$, medium effect; and $|r| > .30$, large effect. These are slight modifications of Cohen's original values. Others have noted that small correlations can indicate theoretically, pragmatically, or clinically important effects that are essentially hidden by their low magnitudes. Squaring an already small correlation coefficient only exacerbates the problem. Except when equal to -1 or $+1$, r^2 will necessarily be closer to zero than *r*, thus subjectively indicating weaker evidence for an effect (recall that for H_0, $\rho = 0$). An infamous example pertains to the correlation between heart

attack occurrence and the ingestion of aspirin (Howell, 2007). In a sample of 22,071 men, these two variables were correlated only .034, with an r^2 value of .001; however, examination of the raw frequencies revealed that nearly half as many men who ingested aspirin suffered heart attacks ($n = 104$) compared to men who ingested a sugar pill placebo and suffered heart attacks ($n = 189$). For those who did not suffer heart attacks, 10,933 ingested the aspirin, whereas 10,845 ingested the placebo. The computed odds ratio was nearly 2:1 in favor of the aspirin preventing a heart attack, and this example shows that such an index of association may be more informative than the correlation or squared correlation.

It is important to note, however, that the odds ratio is not impervious to criticism. First, it is likely to be misinterpreted by laypersons as a probability that can be applied to individual cases—for example, when a man reasons, "If I take aspirin on a daily basis, I will be half as likely to suffer a heart attack." The odds ratio is not a probability, and even if it were, probabilities cannot be meaningfully applied to a single event.[13] More men were found to suffer heart attacks if they ingested the placebo, but each man either suffered a heart attack or remained healthy. Second, odds ratios may subjectively inflate the magnitude of an effect. Stating "men who ingested aspirin were nearly half as likely to suffer from a heart attack compared to men who ingested a placebo" sounds like a dramatic effect, but it masks the importance of the context of the entire sample. The aspirin study was immense, comprising 22,071 participants, only 293 of whom (1.33%) suffered a heart attack. The statistically significant effect hinges on approximately 80 men—only 0.39% of the sample. If 40 men who suffered

[13] A probability is here defined as the ratio of the number of independent elements in an event space (E) to the number of independent elements in a sample space (S), $p(E) = n(E)/n(S)$. Multiple elements are needed in S to compute the probability for E. This is the standard frequency definition of probability that is appropriate for interpreting the heart attack research. Pointing out that a probability cannot be applied to a single observation in no way denigrates the value of probabilities in psychological research, clinical practice, or science. It is simply a matter of applying concepts carefully and accurately. Defending the application of probabilities to single cases, Paul Meehl (1973) wrote, "The vulgar error is the cliché that 'we aren't dealing with groups, we are dealing with this individual case.' It is doubtful that one can profitably debate this cliché in a case conference, since anyone who puts it quite this way is not educable in 10 minutes" (p. 234). Despite such sharp language, the fact remains that a frequency probability cannot be meaningfully applied to an individual event. The man doing so with regard to ingesting aspirin will either suffer a heart attack or he will not. If the aspirin sets into motion the beneficial causal forces, whatever they may be, then the positive effect will necessarily follow.

heart attacks were moved from the placebo to the aspirin group, and 40 men who did not suffer heart attacks were moved from the aspirin to the placebo group, then the effect would not be statistically significant and the odds ratio would be approximately 1:1. In fact, when analyzed with the Observation Oriented Modeling software, this point is clarified by the percent correct classification (PCC) index, which is only 50.39%. The *c* value is still quite low (<.01, 100 trials), but the effect (heart attack) could not be brought into high conformity with the cause (aspirin). The median classification strength index values were also modest, .71, for both correctly and incorrectly classified observations. Thinking in terms of accurately classifying observations on the basis of causal mechanisms, these analyses suggest a much more modest interpretation of the overall results. The incidence of heart attack was relatively low (1.33%), and quite a few men who consumed the aspirin (35%) still suffered a heart attack.

Returning to the narcissism study discussed previously, another limitation of the odds ratio is that it can only be computed for pairs of dichotomous variables. Recall that the ratings were made on 7-point scales, and the correlation between the judge's ratings and self-ratings was .30. Using Hemphill's cut-points, this result represents the upper bound for a medium effect, but how accurate is the judge? In a similar study of narcissism prediction, correlations of .25 and lower were described as indicating "considerable accuracy," as "sizable considering how little information was available," and as supporting the "remarkable accuracy of snap judgments of narcissism" (Vazire, Naumann, Rentfrow, & Gosling, 2008, pp. 1443, 1445, and 1446, respectively). Such descriptions are made in the context of research in which ratings are aggregated across numerous judges and targets and then evaluated in the Pearsonian–Fisherian tradition. In this regard, these descriptions are entirely normative and do not address accuracy per se. From the perspective of observation oriented modeling, the starting point for analysis is the set of observations that have been ordered into seven units. The judge's ratings are considered to be the effect in that they should conform to the narcissism of the people in the photographs, as indicated by their self-ratings. Ideally, an integrated model justifying the use of a 7-point scale and explaining how narcissism is both encoded into a person's physical appearance and then decoded by the judge would be constructed prior to the study. Nonetheless, analyzing the observations in the Observation Oriented modeling software reveals a modest degree of accuracy for the judge, as shown in the multigram in Figure 7.8. The PCC index was 40%, because the judge's ratings matched the self-reported ratings for 20 of the 50

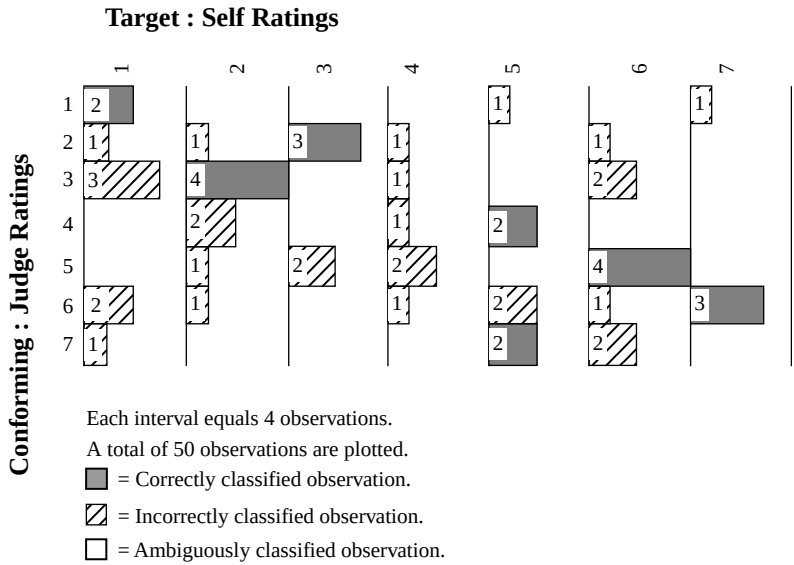
Target : Self Ratings
Conforming : Judge Ratings
Each interval equals 4 observations.
A total of 50 observations are plotted.
= Correctly classified observation.
= Incorrectly classified observation.
= Ambiguously classified observation.

Figure 7.8 Multigram for narcissism ratings on 7-point scales

photographed persons, and the *c* value was low, .06 (1000 trials). As can be seen in the multigram, the accurate judgments seemed to be clustered above and below the midpoint of the scale. Indeed, trichotomizing the scale (1–3, 4, 5–7) yielded an even higher PCC index (58%) and lower *c* value (.03) (Figure 7.9). Such a modification, however, suggests an entirely different understanding of narcissism judgments. Most psychologists employ multipoint rating scales on the basis that the quality being assessed is structured as a continuous quantity in nature. The rating scales are considered to

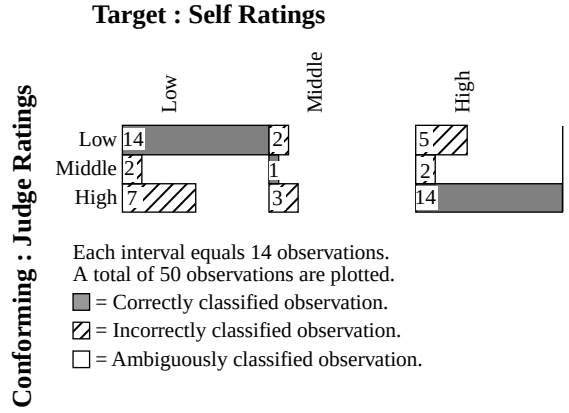

Figure 7.9 Multigram for narcissism ratings on trichotomized scales

represent crude but effective approximations of these continuous structures, and the trichotomized scale would contradict this reasoning. Much more is said about continuous quantities in Chapter 8. The point here—as it was for the in-group/out-group study discussed previously—is that in observation oriented modeling, the observations can be regrouped into different units, but such changes should be made in the context of an integrated model. The goal is to explain the causal mechanisms underlying the accuracy of judgments regarding narcissism. The use of a 7-point scale must ultimately be justified in the context of such causal explanations, and any index of "effect size" must also be wedded to the ways in which the observations are organized into units.

In observation oriented modeling, the PCC index replaces all of the aggregate-based effect sizes from the Pearsonian–Fisherian tradition, and it brings accuracy to the forefront of any analysis in a way that is readily understandable by both scientists and laypersons. Describing correlation coefficients of approximately .25 as indicating "remarkable accuracy" runs the risk of essentially redefining the word "accuracy" and entrapping psychologists in a world of esoteric language that is entirely foreign to other scientists.[14] This is symptomatic of the Comtian positivism underlying the traditional approach. By returning to the observations, the current judgments can be precisely described in the units in which they were made. The judge accurately matched the self-reported units of observation for 20 of the 50 photographs (40%). Discussions of whether or not this level of accuracy is impressive can subsequently proceed without the ambiguity inherent in aggregate measures of effect size. Observation oriented modeling thus offers a distinct remedy to the many ills identified by the APA task force, but it requires a complete break from the positivism of variable-based modeling and a move toward the moderate realism of discerning patterns in observations.

[14] In most studies of interpersonal accuracy, a number of judges evaluate numerous target stimuli. The ratings are then averaged for each judge and sometimes averaged across judges as well. Such aggregation is inappropriate from the perspective of observation oriented modeling because any integrated model of personality judgment will be designed to explain each judge's level of accuracy. Such aggregation has also facilitated the unfortunate redefinition of the term "accuracy," at least in the domain of personality psychology. For instance, Naumann, Vazire, Rentfrow, and Gosling (2009) state, "Our criterion measure [of accuracy] was based on the aggregate of the targets' self-ratings and their informants' ratings of them, a criterion now widely regarded as the gold standard in personality research" (p. 1664). They go on to assess accuracy using correlation coefficients and interpreted their magnitude using conventions.

CHAPTER 8

Measurement and Additive Structures

Contents

Introduction 143
Continuous Quantitative Measurement 145
Measurement as Knowing Quantity 151
Observation Oriented Modeling 158
Additive Models 166
Measurement Error 169
Latent Variables 175

INTRODUCTION

> *Weights and measures may be ranked among the necessaries of life to every individual of human society. They enter into the economical arrangements and daily concerns of every family. They are necessary to every occupation of human industry.*
>
> —***John Quincy Adams (as quoted in Crowley, 1996, p. 65)***

These words, penned by John Quincy Adams in a report to Congress in 1821, are no doubt especially true for the modern scientist, whose very relationship with nature is framed by formal methods of measurement and controlled observation. For the psychologist, too, measurement and observation are defining features of daily life in both the field and the laboratory. With particular regard to measurement, almost every psychologist is trained to understand and apply the four scales of measurement (nominal, ordinal, interval, and ratio) in his or her research. Stanley Stevens first defined the four scales of measurement in 1946, and since then they have become a staple in the lexicon of psychologists on a par with the ubiquity of null hypothesis significance testing (NHST). In fact, the four scales of measurement have become intimately bound to statistical modeling and significance testing because they are often used to guide the choice of analysis for a given study. When dealing with interval or ratio scales, for instance, psychologists understand that a wide variety of statistically

Observation Oriented Modeling
ISBN 978-0-12-385194-9, Doi:10.1016/B978-0-12-385194-9.10008-8
© 2011 Elsevier Inc.
All rights reserved.

powerful methods can be employed, such as Pearson's correlation, *t* tests, or analysis of variance. When dealing with nominal or ordinal scales, less powerful methods must be used, such as Spearman's correlation, Pearson's chi-square, or the Mann–Whitney *U* test. The popularity and pragmatic utility of Stevens' four scales of measurement, however, have overshadowed serious questions regarding the definition of measurement on which they are built, and because it is central to natural science and the day-to-day work of psychologists, any misunderstanding of measurement would prove nefarious.

The foundational concept of measurement is therefore discussed in this chapter, and it is demonstrated that psychologists have in fact misunderstood the traditional meaning of the term. It is also argued that the ill effects of this misconstrual have been as far-reaching as those of NHST, as discussed in the previous two chapters. Rooted in an idealistic philosophy, Stevens' understanding of measurement greatly facilitated the assimilation of the four scales of measurement into the Pearsonian–Fisherian tradition. Once assimilated, they bolstered references to the various social sciences as sciences, per se, while also playing the pragmatic role as guideposts for selecting the appropriate statistical analysis for a given hypothesis and set of variables. As noted by Joel Michell (1999), however, in their efforts to secure and defend the label "science," psychologists skirted a fundamental scientific question: Are attributes such as intelligence, introversion, depression, conscientiousness, need to achieve, as well as countless others structured as continuous quantities? A survey of the most popular statistical analysis procedures currently in use, including *t* tests, analysis of variance, multiple regression, and factor analysis—all of which presuppose continuous, quantitative variables—clearly demonstrates a widespread belief that an affirmative answer has been given to this question. To the contrary, however, no attribute in psychology has been convincingly demonstrated to possess properties consistent with continuous, quantitative structure, and there have been few attempts to even perform the necessary experiments and analytical work.[1]

[1] The most succinct claim that no psychological attribute has ever truly been measured was made by Günter Trendler (2009): "Notice that in this sense [demonstrated continuous quantitative structure] no psychological attribute has ever been measured" (p. 582). In Chapter 8 of his book, Michell (1999) cites a handful of studies in which attempts were made to use additive conjoint measurement to establish initial evidence for continuous quantities.

From the standpoint of observation oriented modeling, this is an untenable state of affairs because the assumption of continuous quantitative structure must be tested if integrated models such as those discussed in previous chapters are to be successfully developed. For instance, how can intelligence, introversion, or any other attribute be posited as an efficient cause or an effect in an integrated model if it is itself not properly understood? The history behind Stevens' four scales of measurement is thus reviewed in this chapter, and an explanation of the continuous quantitative structure assumption is provided. The four scales of measurement are then discarded, and an alternative approach based on observation oriented modeling is presented and demonstrated via analysis of several data sets.

CONTINUOUS QUANTITATIVE MEASUREMENT

Imagine a research psychologist who brings students into her lab to participate in an investigation of personality, depression, and human intelligence. She first asks each student to indicate gender as "M" or "F," which she codes as "1" and "2," respectively, and then to complete a personality inventory that allows her to rank 12 psychogenic needs (e.g., autonomy, dominance, and harm avoidance) of the student from greatest (1) to least (12). Next, each student completes a depression questionnaire that yields scores ranging from 0 (low depression) to 63 (high depression). Finally, each student completes an intelligence test that assesses how quickly the student makes correct judgments regarding novel logic problems. Scores on the intelligence test are recorded in seconds, with high values indicating slower reaction times (lower intelligence) and low values indicating faster reaction times (higher intelligence). Based on Stevens' perspective, the researcher has made four measurements corresponding to the nominal, ordinal, interval, and ratio scales of measurement, respectively. Each is considered as measurement because the psychologist has assigned "numerals to objects or events according to rule," which is the definition of measurement put forth by Stevens and reiterated by authors of virtually every contemporary research handbook and statistics textbook for psychology (Stevens, 1946, p. 677).

Counterbalancing the ubiquity of the four scales of measurement, however, is the almost universal ignorance of the history surrounding their creation. As adroitly researched by Michell (1999), it is not widely known that the British Association for the Advancement of Science appointed a committee in 1932 to investigate the claims that psychophysics, as embodied in the research of Gustav Fechner, Louis Thurstone, and Stanley

Stevens, for example, had established the continuous quantitative structure of sense impressions such as the perception of color, brightness, or sound. According to Michell, establishing continuous quantitative structure was considered as tantamount to measurement, which at the time was itself understood to be "the discovery or estimation of the ratio of some magnitude of a quantitative attribute to a unit (a unit being, in principle, and magnitude of the same quantitative attribute)" (p. 15). With this definition in mind, the committee—called the Ferguson Committee—reviewed the extant research and concluded in 1940 that psychologists had not successfully measured the different sensory attributes.

Given the understood centrality of measurement to science, particularly in physics, it is not surprising that the committee's conclusion was threatening to psychologists, who could have responded in at least one of two ways.[2] First, they could have sought to establish the continuous quantitative structure of sensory perceptions utilizing the definition of measurement set forth in 1901 by Otto Hölder.[3] Fechner's "just noticeable differences" and Stevens' sone scale of loudness fell short of the mark, but further efforts could have been made. Second, psychologists could have charged that the definition of measurement was essentially inaccurate, or at least incomplete. This was the route chosen by Stevens, who adopted an operationist view and defined measurement as the assignment of numbers to objects or events according to rules. This definition had the appearance of being more general than the definition of measurement employed by the Ferguson Committee because the latter was ostensibly included in the former as ratio (and interval, to a lesser degree) scaling. It was also decidedly idealistic, placing emphasis on the rules one follows to assign numbers to observations rather than on determining the natures of the attributes (e.g., perceptions) under investigation:

> *The fact that numerals can be assigned under different rules leads to different kinds of scales and different kinds of measurement. The problem then becomes that of making explicit (a) the various rules for the assignment of numerals, (b) the mathematical properties (or group structure) of the resulting scales, and (c) the statistical operations applicable to measurements made with each type of scale.*
>
> ***(Stevens, 1946, p. 677)***

2 See Michell (1997). Commentaries from five authors are published with Michell's paper, as is his response. The commentaries and Michell's response are highly instructive and elucidate a number of the philosophical and technical issues involved in measurement.

3 Michell presents Hölder's definition in Chapter 3 of his 1999 book. For an English translation of Hölder's 1901 paper, see Michell and Ernst (1996a,b).

With his last point, Stevens wedded his scales of measurement to the Pearsonian–Fisherian approach, which had already taken root in psychology, thus virtually guaranteeing their widespread adoption by research psychologists and other social scientists. His four scales of measurement also helped to insulate psychology from further criticism that it was not a true science based on measurement.

From Michell's perspective, the long-term consequences of this shift in the definition of measurement have been decidedly negative, creating deep confusion among researchers and ultimately stunting the growth of psychology. Paul Barrett (2008) offers a more direct critique:

> *The real-world consequences of this systematic aversion to properly considering the presumed status of a psychological variable is that our journals are now filled with studies that are largely trivial exemplars of mostly inaccurate explanations of phenomena.... Sophisticated statistical models are now used to produce results that seem to have little real-world practical or even scientific consequence.*
>
> ***(pp. 79–80)***

Returning to the research psychologist discussed previously, what set of phenomena is she attempting to explain in her study? Her answer to this question is predicated on her understanding of the natures of gender, psychogenic needs (personality), depression, and intelligence, which brings her face-to-face with the historic debate over measurement. From Stevens' perspective, she has only to examine the rules she followed to classify her measurement of the attributes as nominal, ordinal, interval, or ratio. For instance, she followed a simple rule of assigning "1" to those respondents who circled "M" on the demographics questionnaire and assigning "2" to those respondents who circled "F." The numbers here represent mutually exclusive categories, or nominal measurement. Following a different set of rules, she could easily have ranked the participants from masculine to feminine on an assumed continuous dimension. By following Stevens, however, she has sidestepped a fundamental scientific question—perhaps the fundamental scientific question—regarding her study: Are gender, psychogenic needs, depression, and intelligence structured as continuous quantities? At first glance, the psychologist might argue that depression and intelligence possess such structure because they are classified as interval or ratio scales. A closer look, however, reveals that she has not measured any of her four attributes according to the definition of measurement understood by the Ferguson Committee and formalized by Hölder.

The first step toward understanding why this is so involves turning away from the operationalism and idealism inherent in Stevens'

definition. Consider, for example, the measurement of temperature. Discussions of the four scales of measurement often offer temperature as proof of their efficacy and validity. After all, temperature can be scaled using interval scales (Celsius and Fahrenheit) or a ratio scale (Kelvin). Measuring is the assignment of numbers to objects or events via rules. The rules used to assign the numbers are principal, and the actual nature of temperature is secondary and perhaps even unknowable (recall Pearson's "shadow table" discussed in Chapter 5). In short, the focus is on the knower rather than the things known. The classic definition of measurement, in contrast, rests on a realist assumption that things in nature are the way they are due to inherent features of their informed matter. Consequently, and ironically, the fact that temperature can be measured on scales that differ by constant orders of magnitude (e.g., $°C = (°F - 32) \times 5/9$) actually supports the classic view. There is something inherent to temperature that makes it scale invariant, and through careful, controlled experimentation, scientists have convincingly demonstrated its continuous quantitative nature.

The second step is to realize that only continua can truly be measured according to the axioms of measurement put forth by Hölder. The dimensional aspects of material things—referred to as extensive quantities by Aristotle—as well as their movement through space and time are the most familiar examples of continua. A pencil, for example, is a manifold whole (continuum) whose length can be conceptually divided into centimeters, millimeters, or even angstroms. In principle, the pencil's length can continually be conceptually divided into increasingly smaller units of equal magnitude that approach a limit of zero. The psychologist must therefore determine if the attributes she is studying can be demonstrably divided in an equivalent way, and her demonstrations must have both a logical and an empirical basis. Focusing on intelligence, which is the most likely candidate for true measurement in her study, can the psychologist claim to have measured intelligence as reaction time to novel stimuli? After all, reaction time can be divided into equal intervals (e.g., milliseconds) of increasingly smaller sizes. Measurement from the traditional view, however, is more than simply assigning numbers to objects or events via rules; in other words, it is more than a conceptual or purely subjective process. The psychologist must also demonstrate empirically that the observed reaction times can be divided into equal intervals or, specifically, she must demonstrate the additivity of the reaction times themselves. As Michell (1999) explains,

Let a, b, c, …, etc. be any lengths in the range of all lengths. Then the fact that length is additive is just the fact the following four conditions obtain.

1. *For any lengths, a, and b, one and only one of the following is true:*
   (i) $a = b$;
   (ii) there exists c such that $a = b + c$;
   (iii) there exists c such that $b = a + c$.
2. *For any lengths a and b, $a + b > a$.*
3. *For any lengths a and b, $a + b = b + a$.*
4. *For any lengths a, b and c, $a + (b + c) = (a + b) + c$.*

(pp. 48-49)

This means that if the nature of intelligence is structured as a continuous quantity, then it should demonstrate additive properties.

How might the psychologist demonstrate additivity? Suppose she started with six problems of increasing difficulty (P_1, P_2, etc.) that she wrote and then asked participants to solve. As they solved the problems, she recorded their reaction times and considered longer reaction times to indicate lower intelligence. Consider two participants' data in which each unit of measure is presented as '–' or as a number in a manner similar to a number line:

```
Participant #1
                  P1 P2          P3               P4  P5          P6
   Time(sec) ----3--5---------10------------17--19--------24----

Participant #2
                    P1 P2          P3               P4  P5          P6
   Time(sec) ------4--6---------11------------18--20--------25----
```

Note that Participant 2 is overall slower than Participant 1. These data demonstrate additive structure because for both participants, first, $P_2 - P_1 = P_5 - P_4$ (5 - 3 = 19 - 17) and $P_3 - P_2 = P_6 - P_5$ (10 - 5 = 24 - 19); and second, $P_3 - P_1 = P_6 - P_4$ (10 - 3 = 24 - 17). When thinking of additive structure, it is helpful to consider laying metal rods of equal length end to end. The two small rods for the first three items ($P_2 - P_1$ and $P_3 - P_2$) are equal in length to the two small rods for the second set of three items ($P_5 - P_4$ and $P_6 - P_5$). The combined length of each of these sets of smaller rods is also equal [i.e., $(P_2 - P_1) + (P_3 - P_2) = (P_5 - P_4) + (P_6 - P_5)$]. If the psychologist were to find this same structure for her six items in the responses of numerous people, then she would have initial evidence that intelligence itself is structured as a continuous quantity. Having passed this first hurdle, suppose she then wrote, based on her understanding of intelligence as speed of processing, additional items located precisely between P_2 and P_3 or between P_3 and P_4 on the number lines presented previously. If she succeeded in writing such items and obtaining data such as those presented previously across numerous

participants, she will have made great strides toward establishing intelligence as a continuous, quantitative attribute.

For comparative purposes, suppose the observed reaction times for the two participants discussed previously were instead as follows:

```
Participant #1
                  P1 P2       P3      P4      P5  P6
   Time(sec) ----3--5------9------13----16--18—

Participant #2
                            P1          P2  P3 P4 P5              P6
   Time(sec) --------------8--------13-14-15-16------------23--
```

Although the ranking of the reaction times is preserved (i.e., item 1 is solved quickest and item 6 is solved slowest), there is no clear additive structure in the data when comparing the two participants. In this case, although the psychologist used reaction time to "measure" intelligence, the results do not demonstrate additive structure and therefore appear not to support the hypothesis of continuous quantitative structure. As one could imagine with genuine data, the picture would likely be even murkier because participants would differ in their orderings of the items solved most quickly. For example, a participant's reaction times might be as follows:

```
                     P2 P6    P4     P1  P3        P5
Time(sec) ----3--5----8----11-13------17--
```

The ordering of the items for this participant did not remotely match the expectations of the psychologist.

Establishing additivity is therefore an important part of demonstrating the continuous quantitative structure of an attribute such as intelligence. In summarizing Hölder's theory of measurement, however, Michell (1999, pp. 51–53) presents a total of seven conditions that must be met and readily admits that the road toward establishing continuous quantitative structure is long and arduous.[4] The psychologist must nonetheless travel this road if her theory commits her to positing intelligence and the other attributes as continuous quantities in nature. However, can she arrive at her destination? In other words, can she devise methods for observing each attribute in a manner that fulfills the requirements of Hölder's theory of measurement? Answering this question first requires a broader philosophical conceptualization of the term measurement.

[4] See also Michell (2008, p. 16). Volume 6, 2008, of the journal *Measurement* is a special issue devoted to Michell's critique. Commentaries with rebuttals provide interesting insights into measurement, item response theory, latent variables, and other related topics.

MEASUREMENT AS KNOWING QUANTITY

Philosophically defined by Aristotle, measurement is an act of knowing; specifically, it is an act of knowing the quantity of some thing or things in nature. Knowing from the perspective of Aristotle entails a subject–object relationship in which the knower (subject) becomes one with the known (object) in an intentional and immaterial way. In other words, the knower comes to possess the form of the thing known without its matter. Although many find this a difficult teaching to accept, it leaves little doubt regarding the thoroughgoing nature of Aristotle's realism. More pertinently, Aristotle's definition implies that all measurement presupposes unity. This can be understood by first realizing that formal cause delineates more than simply the shape of some known thing. When speaking of substantial form, in particular, reference is made to the "whatness" of a thing or that which gives unity to matter and makes a thing to be what it is rather than something else. Knowing an indivisible *one* is thus the essence of measuring, which is why Aristotle and Aquinas (in his commentary on Aristotle's *Metaphysics*) regard "one" as a principle of both numbering and measuring. As explained by Charles Crowley (1996),

> *Initially the notion of measure (*ratio mensurae*) is derived from number to other quantities, namely, that just as one, which is the principle of number, is indivisible; so in all the other genera of quantity some "indivisible one" is a measure and is a principle (of knowing the quantity of a thing, or of measuring a thing).*
>
> ***(p. 38)***[5]

In the simple act of counting apples, for instance, a person must first know an apple as a unified, existing thing composed of matter and form. The person can then understand the plurality, magnitude, and intensity of apples through measurement. With regard to plurality, the person considers the apples as discrete, countable units in knowing their quantity—for example, in counting the number of apples in a barrel. With regard to magnitude, the person considers an individual apple as a continuum extended in space whose length, width, and depth can be divided (in mind) into discrete indivisible units. In this way, the person can measure an apple's extensive

[5] See also Redpath (2010): "According to Aristotle, a measure is the means by which we know a thing's quantity. And quantity is that by which we know substance. That is, a measure is a unit, number, or limit. Aristotle adds that we first derive the notions of measure and order from the genus of quantity. From this we analogously transfer these notions to other genera. In a way, unity and quantity are the means by and through which we even know substance, quality, everything" (p. 8).

quantity using conventional units such as inches or centimeters. What of intensity? The person has thus far measured aspects involving the extended matter of apples, but there are other properties of the apples that may be measured, such as a particular apple's weight or temperature. How are these measurements to be made?

Crowley (1996) quotes Aquinas as stating,

> *Quantity is twofold, one which is called* molis *quantity, or dimensive quantity, which is found only in bodies.… The other is* quantitas virtutis*, which is taken according to the perfection of some nature or form. This quantity is designated insofar as something is said to be more (*magis*) or less (*minus*) hot, insofar as it is more perfectly or less perfectly hot.*
>
> ***(p. 28)***

The quality "heaviness" can be known by holding an apple in hand, and as Crowley points out, it is the task of the natural scientist to analogically transfer the predicates of dimensive quantity to this quality if it is to be considered as a "quantitative quality" or *quantitas virtutis*. In ancient times, this was done via the comparison of two objects on a two-pan balance, and the objects were considered as equal in weight when the difference between the heights of their pans was judged to be equal to zero. A trip to the doctor's office shows this "primitive" method of measurement is still highly useful. As the patient stands on the balance's platform, the nurse moves a series of counterweights until the difference between the balance's pointer and the midpoint (zero point) on the scale is equal to zero. The nurse then reads and sums the numbers on the arm of the balance corresponding to the positions of the counterweights. What the nurse has accomplished here, according to Crowley, is the analogical transfer of a well-known dimensive quantity (length) to a quality,

> *and the notion of equality … is now attributed to the quality heaviness, enabling that quality to be measured as a* quantitas virtutis*, and to be defined accordingly. Hence, weight is measured heaviness, which is now a "quantity"—whether it be measured by the balance scale, or by the "force of attraction" of the Earth to a body, with or by a coiled spring hook scale.*
>
> ***(p. 29)***

The "hotness" of an apple can be measured via the same principle, namely by examining the height of mercury in a thin, calibrated tube. Although these examples are fairly straightforward, Crowley offers a more thorough and impressive exposition of this principle with regard to the measurement of other qualities central to natural science, and physics in particular (e.g., electric current, force, and luminosity), using the International System of Units.

This admittedly brief philosophical analysis is significant for psychologists in a number of ways. First, it provides the basis for distinguishing between (1) measurement as counting and (2) measurement of magnitude in accord with the previously discussed definitions regarding continuous quantitative structure. The importance of this distinction can be understood by imagining a physiologist who quantifies cortisol concentration in human saliva in nanomoles per liter. Ostensibly, the physiologist has measured cortisol levels in accord with the measurement theory of Hölder; however, closer examination reveals that instead the physiologist has simply counted (or estimated the number of) discrete units of a chemical compound that cannot be subdivided further in any meaningful way, consequently violating the requirement of continuity. A human factors psychologist may similarly count the number of tables in restaurants and report the results in "tables per square foot" units. Both researchers can be said to have measured cortisol or tables, respectively, but it is measuring (i.e., knowing quantity) as counting, not measuring as the conceptual division of a continuum into discrete units.

A second implication is that psychologists must acknowledge the critiques of measurement by Michell, Barrett, and others as valid and crucial. In other words, psychologists cannot afford to simply assume continuous quantitative structure of the attributes they wish to measure; rather, they must test this scientific question as a matter of necessity. Crowley's analysis shows clearly how physical scientists have analogously transferred measurements of dimensive quantities to intensive quantities (or quantitative qualities, *quantitas virtutis*) and no one doubts their success. Psychologists must similarly demonstrate the continuous quantitative structure of the attributes (variables and theoretical constructs) they study as well. Failure to do so makes the development of integrated models such as those discussed in the previous chapters impossible because to create such models, the material and formal causes that are internal to the components of the models must clearly be delineated. A model constructed around intelligence, for example, as a discrete number of counted behaviors is fundamentally different from the same model constructed around intelligence as a continuous quantity. Expressions of efficient and final causes will furthermore be intimately bound to material and formal causes (see Chapter 9), meaning that the issue of continuous quantitative structure is fundamental to every aspect of modeling.

An equally important consideration is that empirical tests of a given integrated model must employ analysis techniques appropriate to the

components of the model. Psychologists are well aware of this principle, but insofar as it involves the use of Stevens' four scales of measurement in the choice of the appropriate statistical analysis. Measurement herein, however, is understood generally as either counting or determining the magnitude of an extensive dimension or intensive quality. Aristotelian in origin, it is founded on philosophical realism and therefore fundamentally incompatible with Stevens' four scales of measurement. As noted previously, Stevens created a primarily operationist and idealistic (specifically, subjective) formulation of measurement that, according to Michell, also contradicted the philosophical realism (albeit not necessarily Aristotelian) underlying the classic view of measurement employed by the Ferguson Committee and formalized by Hölder.[6] The bottom line is that rather than following rules for assigning numbers to objects or events in order to pick the appropriate statistical test for a variable-based model, the psychologist should instead focus on developing an integrated model with careful attention to its components (qualities, attributes, structures, processes, etc.). If a given component is posited as a continuous quantity, then the arduous road of substantiating this claim must be taken; otherwise, the component must be treated as counted discrete units or counted ordinal judgments.

When constructing integrated models, psychologists should not shy away from discrete quantities, considering them to be somehow inferior to continuous quantities. Indeed, continuous measurement presupposes the establishment of discrete countable units that possess certain features, such as equality, uniformity, and regularity (invariance). Discrete quantities can also convey information about formal causes, such as when chemical compounds are denoted as ratios of elements—for instance, H_2O, CO_2, and $C_{21}H_{30}O_5$ (cortisol).[7] More important, finite mathematics, which presupposes discrete measurement, has grown by leaps and bounds in the past 100 years, particularly with the advent and widespread accessibility of the digital computer. Set theory, graph theory, Boolean algebra, combinatorics, and logical analysis (see Chapter 9) represent a subset of a larger number of topics covered in finite mathematics. Computer models of

[6] For a discussion and defense of philosophical realism, more generally considered, see Harré (1987).

[7] Crowley (1996) states, "It can be said that the essence of compounds chemically is number, which is the formal cause of chemical compounds insofar as so much of one chemical element (e.g., hydrogen) is to so much of another chemical element (e.g., oxygen), to form the chemical compound (e.g., water), and expressed in the formula H_2O, which is not number but a ratio of numbers" (p. 58).

self-organizing, self-replicating, or dynamic systems are also often constructed on the basis of discrete events or discrete units of observation, such as cellular automata.[8] Psychologists concerned primarily with discrete measurement therefore have at their disposal a diverse and powerful set of modeling tools, and they need not worry about losing their status as per se scientists. The whole of natural science is composed of more than measuring continuous quantities as surely as the natural world is populated by things that cannot solely be conceptualized in terms of dimensive quantities. Scientists are obligated to understand this fact and conform their models to reality rather than squeeze nature into disfiguring molds.

A brief statement about ranks should be made here because they were assigned their own category by Stevens. Given the previous discussion, ranks are not to be considered as a separate form of measurement, and in modeling they will likely result from secondary judgments of counted units or magnitudes (e.g., ranking the number of cortisol compounds from greatest to least). Ranks might also be found in primary judgments of "more," "less," "greatest," and "least" when two or more things are compared (e.g., "Who is most intelligent, Einstein, Newton, or Galileo?"). These latter judgments involve a different predicating process than those directed toward quantities; for instance, stating that two people are *the same* regarding their heaviness is different than stating their measured weights are *equal* (e.g., 185 lbs = 185 lbs). The predicate "equal" only applies to quantities, whereas "same" applies more generally to qualities. In addition to the quantitative qualities (*quantitas virtutis*) discussed previously, Aristotle enumerated different types of qualities (e.g., habits, powers, and shape), some of which could be predicated as "more" or "less" (Aristotle's *Organon*, Categories, Chapter 8). Alternatives or additions to Stevens' four scales of measurement include options that involve particular forms of predicating qualities (Chrisman, 1998; Velleman & Wilkinson, 1993). For instance, with a grades scale, observations are recorded as ordered labels, such as Freshman, Sophomore, Junior, and Senior. Other proposed scales involve proportions, amounts (non-negative real numbers), and geometric angles, but all can ostensibly be understood through counting or the determination of magnitude (as continuous quantitative structure). Each of the grades scale units is

[8] Stephen Wolfram's *A New Kind of Science* (2002) offers an overview of modeling with cellular automata. This new kind of science, however, is not so new and can be traced back to the late 1960s and Konrad Zuse (1967, 1969). John Conway's Game of Life popularized cellular automata in the 1970s (see Gardner, 1970).

a countable observation, amounts correspond to counting some base unit (e.g., counting money in pennies), and angles are defined as ratios of lengths.

Returning to the previous question, can the psychologist devise methods for observing each attribute in a manner that fulfills the requirements of Hölder's theory of measurement? Clearly, if her scientific model dictates that intelligence, for example, possesses continuous quantitative structure, she must attempt to answer this question in the affirmative. Her efforts will involve analogously transferring an extensive quantity to the intensive quality of intelligence. As previously made clear, it is not sufficient to simply use a measuring method that yields a well-known magnitude, such as reaction time. She must attempt to demonstrate the additive structure of intelligence using the logic also outlined previously. From Michell's perspective, a more adequate starting place for the psychologist would be the theory of conjoint measurement first elucidated by Duncan Luce and John Tukey in 1964. The details of this theory are not presented here, but it essentially provides the analytical and technical framework necessary for attempting to demonstrate additivity. It has been used with a modest degree of success since its inception, although it has been famously referred to as the "revolution that never happened" (Cliff, 1992). The said revolution was to be one in which psychologists and other social scientists finally tested the assumed continuous quantitative structure of their most highly marketed attributes. Regardless, Michell (1999) admits that successfully applying conjoint measurement theory would not alone provide a convincing argument for true continuous quantitative structure:

> *Conjoint measurement theory is but one conceptual resource amongst an array. Its significance is that it fills a specific, debilitating gap in the quantitative psychologist's methodological armory.... It indicates a place to start, but the journey will only ever be completed in conjunction with advances in substantive areas of psychology and a deeper understanding than we have now of how psychological systems work.*
>
> ***(pp. 207–208)***

The psychologist might therefore, in the context of an integrated theory, begin with additive conjoint measurement. Assuming she is successful, she might then move on to the equally difficult work of experimentally manipulating causes and effects involving the attribute (quality) under investigation. Such work is always necessary for determining the continuous quantitative structure of a *quantitas virtutis*. Previously, it was noted that temperature could be measured analogously as the calibrated height of mercury in a thin tube. This fact, however, depends on the effect of thermal

energy on mercury; namely, increasing levels of thermal energy causes mercury to expand.[9] The tube can be calibrated by carefully controlling different factors (e.g., the diameter of the tube and the quantity of mercury), thereby systematically and precisely manipulating this cause–effect relationship. The psychologist must similarly be able to manipulate the cause and effect relationships involving intelligence to demonstrate its continuous quantitative structure; hence, the critical question becomes, Can she do so?

In answering this question, Günter Trendler (2009) argues that psychologists, and by extension other social scientists, will never have sufficient control over the attributes they wish to measure as continuous quantities. At the level of overt, observable behavior, Trendler views psychological phenomena as simply too unwieldy and beyond the scientist's capacity to precisely control causes and effects or to control unwanted disturbances. Even at the level of neuropsychological functioning, Trendler sees no hope of constructing apparatuses that would in any way permit a researcher to isolate a given attribute (e.g., intelligence and introversion), let alone manipulate it in some systematic and precise manner. Stated in words consistent with the previous discussion, it is impossible for any psychologist or social scientist to ever convincingly (although technically still analogously) transfer a known extensive quantity to intensive qualities of human perception, intelligence, personality, emotional states, etc. In a stunning conclusion, then, Trendler not only states that psychological phenomena are not measurable—in the sense of having demonstrable continuous quantitative structure—but also states that measurement theory (e.g., conjoint measurement theory) is of absolutely no consequence to the future development of psychology—a conclusion presaged by Peter Schönemann 15 years prior (Schönemann, 1994).[10] These conclusions of course extend to phenomena studied by other social scientists. Trendler's recommendation to

[9] These points and examples are taken from William Wallace's *The Modeling of Nature* (1996, p. 242).

[10] Schönemann presages Trendler's conclusions in his provocative book chapter: "It is far from self-evident why the Archimedean axiom should hold in psychology, or in biology, where most phenomena are bounded by physiological constraints. Nor is it self-evident why it should always be possible, or even helpful, to remove interactions as additive conjoint measurement tries [sic] to do. Why should the 'crisp' mathematics of physics apply without change to the fuzzy nature of living things.… None of this is self-evident *a priori*, nor is any of it empirically founded" (p. 158). Schönemann also discusses examples of contrary evidence simply being brushed aside or ignored, and he discusses the various forces that maintain the blindness to this fundamental issue of science.

the psychological researcher discussed previously would thus be to give up the impossible task of demonstrating continuous quantitative structure and to spend her time more wisely developing alternative methods:

> *Other, more suited methods for the domain of psychology must be found. It might therefore be wise to seriously reconsider Johnson's recommendation: "Those data should be measured which can be measured; those which cannot be measured should be treated otherwise. Much remains to be discovered in scientific methodology about valid treatment and adequate and economic description of nonmeasurable facts".*
>
> ***(Johnson, 1936, p. 351, in Trendler, 2009, p. 593)***

OBSERVATION ORIENTED MODELING

From the perspective of observation oriented modeling, psychologists normally work with discrete countable units and ordinal judgments rather than continuous quantities.[11] As discussed previously, this standpoint is supported by the failed efforts of psychologists and other social scientists to convincingly demonstrate the continuous quantitative structure of their most highly marketed qualities (or attributes) and by Trendler's arguments that show convincingly such demonstrations will not be forthcoming. The shift in thinking away from variable-based models in the previous two chapters is therefore supported further, and it can be expanded upon here by comparing bivariate and multiple regression to the alternatives provided by observation oriented modeling.

Most high school students learn the equation for a straight line as $y = mx + b$, where y and x refer to values of the ordinate and abscissa axes of a Cartesian coordinate system, respectively, m is the slope of the line, and b is the y-intercept of the line. Using this function, a person can transform a value of y to x and vice versa if the values for m and b are known. For instance, if $m = 2$, $b = 1$, and $x = 5$, then $y = 11$. If $y = 21$ for the same function, then x can be solved for and found equal to 10.

[11] Additional support for this assertion can be found in a series of articles and letters to the editor in the 1989 volume of *Archives of Physical Medicine and Rehabilitation*. The lead article in the series is by Merbitz, Morris, and Grip (1989). The letters to the editor with replies from Merbitz et al. can be found in the same volume. The well-known measurement specialist, Benjamin Wright also commented on Merbitz et al.'s article (Wright & Linacre, 1989). Wright and Linacre's article essentially advanced the Rasch model underlying item response theory as sufficient for demonstrating interval scaling and continuous quantitative structure. As noted, however, Joel Michell's work has shown that the Rasch model is not sufficient and rather assumes quantitative structure of the attributes attempting to be measured.

It is customary when deriving m and b from genuine data to write the equation for a straight line as $y = a + bx + e$, where y is denoted as the dependent variable, a is the y-intercept, b is the regression weight (slope parameter), x is the independent or predictor variable, and e is a random variate. Given a set of values for y and x, the method of least squares can be used to solve for the y-intercept and slope parameter in the equation. The scatter plot in Figure 8.1 shows the relationship between the boiling point of water (y, in °F) and barometric pressure (x, rescaled by $\log_{10}$) (Weisberg, 1985). The observations form an almost perfect straight line, defined as $y = 110.11x + 49.26$ using the method of least squares. If one were to take any observed value for pressure and transform it to a corresponding value for temperature using the equation, the transformed value would be extremely close to the observed value. Moreover, because both variables are known to possess continuous quantitative structure, any real value of x could be chosen and converted to a corresponding value of y. For example, for $x = 1.35$, $y = 197.91$, and for $x = 1.36$, $y = 199.01$. These particular observations were not made and are hence not plotted in the graph, but they can be held with confidence as closely representing the actual pressure measurements that would be taken because of the accuracy for the existing 31 data points and the known continuous quantitative structure of temperature and barometric pressure. The median absolute difference between the actual and predicted (based on the formula) temperatures is

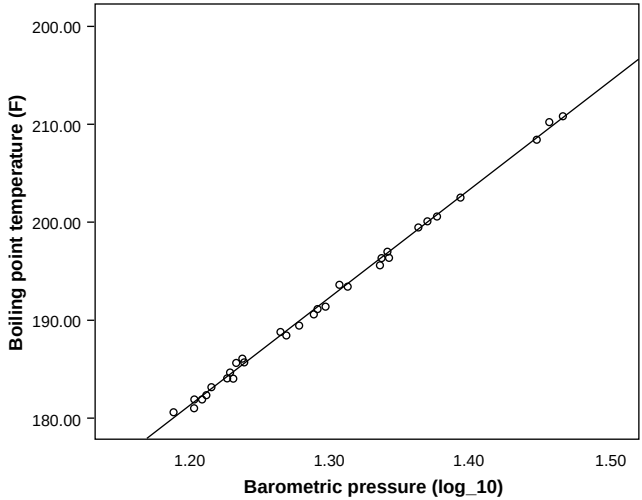

Figure 8.1 Scatter plot of boiling temperature for water and barometric pressure. Data from Weisberg (1985).

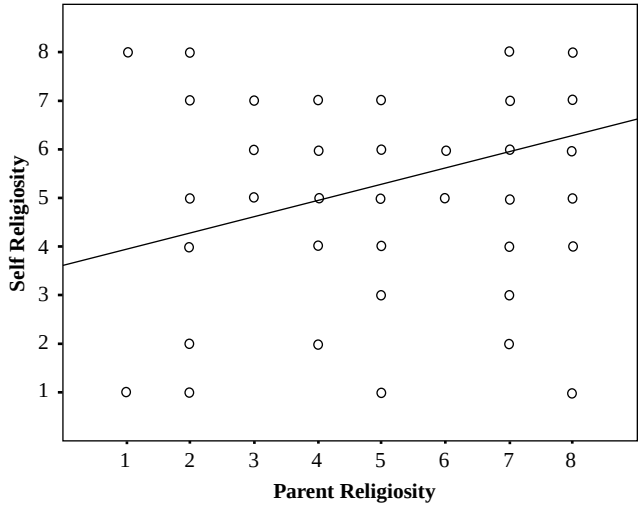

Figure 8.2 Scatter plot of rated self and parent religiosity

only .26 degrees, which is also reflected in a near-perfect multiple R^2 value for the regression analysis ($R^2 = .998$, maximum = 1).

By way of comparison, consider the scatter plot in Figure 8.2, which shows the relationship between two items intended to measure the variables "religiosity" and "parental religiosity," both from the perspective of the respondent:

I am a religious person.

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Definitely False	*False*	*Mostly False*	*More False than True*	*More True than False*	*Mostly True*	*True*	*Definitely True*

My parents are religious people.

Definitely False	*False*	*Mostly False*	*More False than True*	*More True than False*	*Mostly True*	*True*	*Definitely True*

Given the general trend in the pairs of observations plotted for 80 different people, supported by accompanying statistical analysis ($R^2 = .13$, $p < .001$) and a regression line with a positive slope, $y = 3.61 + .33x$, most psychologists would describe the relationship as approximately linear.[12] At the level of individual cases, or hypothetical cases not included in the data set, a change in parental religiosity scores from 1 to 2, for instance, would be interpreted as corresponding to a change in self religiosity scores

12 These data are from Kassing (1997). The wording of the original copyrighted item stems has been altered.

from 3.94 to 4.27. Similarly, a change in parental scores from 1 to 3 would correspond to a change in self scores from 3.94 to 4.60. The change in parental scores from 1 to 3 scale points thus corresponds to twice the amount of predicted change in self scores compared to the 1 to 2 point change in parental religiosity. This simple equality of ratios of differences $[(3 - 1)/(2 - 1) = (4.60 - 3.94)/(4.27 - 3.94)]$ is only computationally and conceptually legitimate, however, if the judgments of religiosity are truly structured as continuous quantities. Clearly, continuity cannot reasonably be assumed for ratings obtained from the two scales, not to mention additivity and Hölder's axioms of measurement. However, very few psychologists would question this analysis or even flinch at the sight of predicted values (3.94, 4.27, and 4.60) in the form of real numbers that can never be observed on the 8-point rating scales. Most psychologists would in fact pay little direct attention to the predicted values at all and would instead focus on the statistical significance of the analyses and the index of effect size, as discussed in the previous two chapters.

This bias not only attests to the ubiquity of the continuity assumption and the damaging blindness it can cause but also reflects a tacit trade-off that has been made in the history of psychology; namely, accuracy has been traded for statistical significance, variable model fit, and the assumption of continuous quantitative measurement.[13] As some have argued, the development of this

[13] This trade-off can be seen in an interesting resistance that sprang up briefly after the introduction of Stevens' four scales of measurement. The claim by some psychometricians and methodologists at that time was that the legitimacy of mathematical operations, and therefore statistical analysis, should not be driven by concerns regarding how data are scaled. In a famous and colorful note to the *American Psychologist*, Frederick Lord (1953) parodied a statistician who defiantly and successfully applied mathematics and sampling techniques to answer a question regarding numbers (nominal data) on football jerseys. He claimed that such operations were fine because "the numbers don't remember where they came from, they always behave just the same way, regardless" (p. 751). Others followed suit, arguing that as long as principles of sampling and underlying assumptions are not violated, statistical analyses can be applied to different types of data. For instance, if the assumptions of normal, independent, and homogeneous errors are met, the p value for multiple regression will be legitimate and the analysis may therefore be applied to ordinal data. Obviously, these arguments depend on purely practical concerns as well as indices of effect size and statistical significance for their force. They leave unanswered the scientific question of continuous quantitative structure and the importance of developing accurate, integrated theories in science. In the end, this Resistance failed to prevent the use of Stevens' four scales of measurement as guides for selecting statistical analyses, but it succeeded in diverting attention away from the serious questions of measurement and theory development (Anderson, 1961; Burke, 1953; Gaito, 1980).

trade has been driven by a "quantitative imperative" or some sort of "physics envy" guiding psychologists (Michell, 1999, Chapter 2). Whatever the case may be, it precludes recognition of the most powerful feature of the function graphed in Figure 8.1—its accuracy. It is true that the linear function relating boiling point of water to barometric pressure is impressive because of the precision with which values can be transformed. Pressure readings out to two, three, or four decimals could be converted to predicted boiling point values. This simply demonstrates the continuous quantitative structure of the two measures. What is more impressive, however, is that the function describes an accurate, not just a numerically precise, mapping of one set of observations to the other. This is of course the more general way of thinking about functions—as relations between domains and ranges whose components may or may not be real values or other quantities. It is with this sense of accuracy—the accuracy of mapping one set of observations to another—that psychologist should primarily be concerned.

As demonstrated in previous chapters, with observation oriented modeling one set of observations is brought into conformity with another, and the accuracy of this transformation stands at the center of the analysis (i.e., the percent correct classification). No assumptions of continuity or additivity are made in the analysis, nor is a particular function assumed. Applying linear regression to the 80 religiosity observations mentioned previously implied that the relationship was expected to be linear rather than curvilinear. Observation oriented modeling does not incorporate a particular function relating two variables but instead works on the basis of the similarities between the deep structures of observations. The plural term "variables" is in fact eschewed in favor of "conforming observations" and "target observations." Considering parental religiosity (at least as perceived by the respondents) as the cause of self-rated religiosity, the goal of the analysis in the Observation Oriented Modeling software is to bring the deep structure of the latter observations into conformity with the deep structure of the former. The judged success of this transformation revolves around accuracy, or simply the tallied number of matches between the conformed (i.e., transformed) and target religiosity observations.

Results of the observation oriented modeling analysis for the same 80 observations in Figure 8.2 reveal an unimpressive degree of accuracy, with the percent correct classification (PCC) equal to 38.75%. Less than half (31) of the 80 participants were classified correctly, with a median classification strength index (CSI) equal to .75. The randomization test, however, yielded an impressively low c value ($<.001$, 1000 trials), suggesting that the pattern

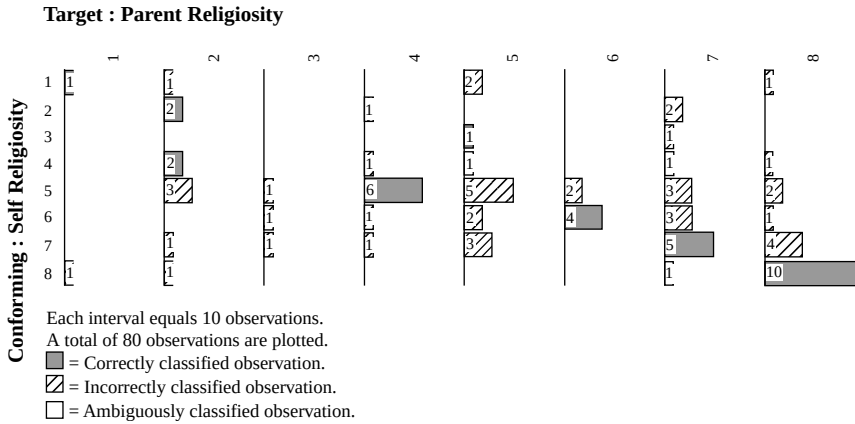
Target : Parent Religiosity
Conforming : Self Religiosity
Each interval equals 10 observations.
A total of 80 observations are plotted.
= Correctly classified observation.
= Incorrectly classified observation.
= Ambiguously classified observation.

Figure 8.3 Observation oriented modeling results for parental and self religiosity ratings.

of observations was nonetheless relatively unique. The multigram in Figure 8.3 shows the configuration of observations with the majority of misclassifications. The overall pattern of the observations does suggest a one-to-one mapping, particularly if the correctly classified observations (the white bars in the figure) are focused on, but the paucity of observations for the lower scale values is an obvious concern.

Two courses of action are usually appropriate in observation oriented modeling when too few observations are available for the different combined units of ordered observation. First, the investigator can return to the laboratory or field and obtain the needed observations. For the current example, the researcher would need to find more persons endorsing the 1, 2, 3, and 4 scale points. If more observations will not be forthcoming, then the investigator must explain why. If it is a matter of insufficient resources, then the existing observations can be grouped into different units. For instance, a sensible and purely post hoc grouping would reduce the observations to "low" (1–4 scale points) and "high" (5–8 scale points) for each of the questions. The observation oriented modeling analysis of these reduced observations for the current 80 individuals yields an improved PCC value of 66.25%, with 53 of the 80 observations classified correctly (c value = .14, 1000 trials). The median CSI value, however, for the correctly (.73) classified observations was slightly lower than the original model and no higher than the median value for the incorrectly classified (.73) observations. The multigram in Figure 8.4 also shows that even with this new grouping, only 8 people

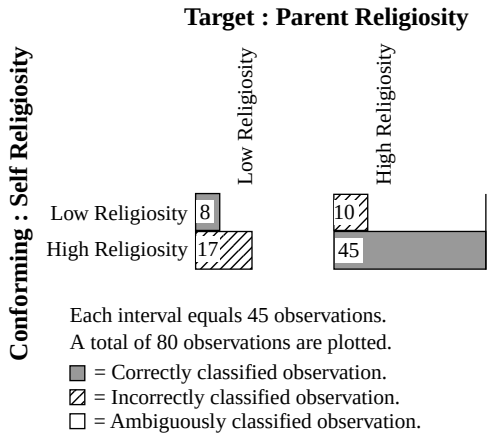

Figure 8.4 Multigram for dichotomized parental and self religiosity ratings

were observed in the low/low group. Moreover, most of the people in the low parental religious group (17 of 25) were also in the high self-rated religious group, contradicting a one-to-one mapping between observations. Still, the correctly classified observations suggest the self-ratings can be modestly conformed (PCC = 66.25%) to the parental ratings in such a way that they correspond in religiosity when the original 8-point scale is dichotomized. However, this conclusion must be considered with several caveats, most notably the paucity of observations for individuals with both low self-reported and parental religiosity.

The regrouping of the scales into dichotomous units points to the second possible course of action when the numbers of observations are too few; namely, changing the original format of the statements in the context of an integrated model or general theory. For instance, the item response format could be changed as follows:

I am a spiritual/religious person : Yes/No/Does not apply

My parents are spiritual/religious people : Yes/No/Does not apply

Conceptually, this format would be consistent with George Kelly's (1955) personal construct theory (PCT), in which human judgment is considered to be fundamentally bipolar in nature. In PCT, each person is metaphorically considered to be a scientist who attempts to make sense of the world through a hierarchically arranged system of bipolar personal constructs (e.g., happy/sad, uplifting/depressing, and free/enslaved). Much as every scientific theory is meant to explain only particular phenomena, each

dichotomous personal construct has a limited range of convenience regarding the things, places, people, situations, etc. to which it is applied. For instance, the construct happy/sad would not normally be applied to furniture, minerals, or planets. A person's "Does not apply" response to one of the previous statements would thus indicate that it lies outside of the range of convenience for the individual's personal constructs. In observation oriented modeling, linearity and continuous quantitative structure are not assumed, and the "Does not apply" response can therefore be treated as one part that makes up the whole item. In other words, it can be included as a unit in the analysis rather than as missing data, as might be done in the Pearsonian–Fisherian tradition.

The response format discussed previously is similar to behavioral checklists used in a wide variety of research and applied settings. For instance, checklists used in many psychological clinics may ask incoming clients to indicate the frequencies of different symptoms experienced during the past week, such as nervousness, sleeplessness, loss of appetite, and confusion. The scaled responses are often summed and compared to various cut-points to determine if the person may be suffering from severe anxiety, depression, or some other clinical syndrome. Checklists are common in medicine as well, but often these checklists are used to determine a logical combination of symptoms that might indicate a particular malady; for instance, a sore throat without fever would indicate a common cold, whereas a sore throat with fever would suggest influenza.

Although checklists may represent simple tallies, they can be very informative and effective. Barrett (2008) discusses one such example, the Violence Risk Assessment Guide, which predicts violent recidivism with an accuracy of 72%, a result that has been replicated in several countries for a checklist that was developed with "no structural equation modeling, no data-model-driven regressions, no 'made-up' latent variables, and, above all, no assumptions as to the 'quantitative' nature of the scale of risk propensity" (p. 82). The cognitive and underlying neuronal processes that explain responses to dichotomously formatted statements may also prove to be more amenable to the development of explanatory models. Robert Schwartz, for instance, developed his states of mind model to explain the frequencies that individuals endorse behavioral symptoms of emotional and psychological distress. According to the model, normal, emotionally healthy adults should assent to positively worded descriptions of themselves with a particular frequency (i.e., .72), which Schwartz likens to a homeostatic set point (Schwartz, 1997; Schwartz, Reynolds, Thase, Frank, & Fasiczka, 2002).

The observation oriented modeler must be willing to pursue these types of models and the methods they entail.

ADDITIVE MODELS

The equation for a straight line can be modified to describe a plane in a three-dimensional or higher order space. Similarly, the basic equation for linear regression can be modified to include multiple independent variables. For example, with two predictors, the equation would be written as $y = a + b_1x_1 + b_2x_2 + e$, and with three predictors it would be written as $y = a + b_1x_1 + b_2x_2 + b_3x_3 + e$. In terms consistent with variable-based models, predictors are added to a regression equation in a bid to explain a greater proportion of variation in the dependent variable (y). They might also be added for theoretical reasons when the dependent variable is considered to be an additive function of several other variables. Consider, for example, a regression equation describing relationship commitment as an additive function of relationship satisfaction, investment, and attractive alternatives:

$$\text{Commitment} = a + b_1 \text{ satisfaction} + b_2 \text{ investment} - b_3 \text{ alternatives} + e.$$

As shown in the equation, higher satisfaction, higher investments, and fewer attractive alternatives predict higher commitment (Rusbult, 1980). Using traditional methods embedded in the Pearsonian–Fisherian tradition, the regression weights and y-intercept terms can be solved for genuine data using the method of least squares. The overall success in relating commitment scores to the combined scores for satisfaction, investments, and alternatives can be quantified with multiple R^2 and tested for statistical significance. The methods applied to the simple linear function presented previously therefore generalize to this and other more complex equations.

Similar models can be tested in observation oriented modeling but in ways that assume orders rather than continuous quantities, although the methods will work with the latter types of observations as well. Building on the previous example, consider the following statement and response format:

I have few unresolved conflicts with my parents.

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Definitely False	*False*	*Mostly False*	*More False than True*	*More True than False*	*Mostly True*	*True*	*Definitely True*

As discussed previously, although the points on the scale are spaced and numbered in equal intervals, the observations obtained from this item

cannot be claimed to possess continuous quantitative structure. They can reasonably be said to represent successive ordinal judgments of "less" or "more," and their deep structure can then be added to the deep structure of the parental religiosity item. Recall from Chapter 2 that deep structure addition is not necessarily equivalent to arithmetic addition because it more generally preserves the ordering of units that constitute an observation. Adding the deep structures of the parental religiosity and conflict response scales yields observations with 15 units (see Chapter 2) that can be labeled 2–16, although they again only represent combined successive ordinal judgments of "less" or "more" for this example. The goal in adding the deep structures of these two sets of observations is that the resulting deep structure will more accurately conform to the self-rating religiosity observations. In other words, the goal is to increase accuracy, not to explain more variance. Ideally, the deep structure addition would be driven by an integrated model, but the simple expectation here is that the causal connection between young adults' religiosity and their parents' religiosity will be enhanced by the extent to which they report having few unresolved conflicts with their parents. As will be seen, this is equivalent to arguing that extending the number of units for the target observations will lead to a more clearly identifiable pattern in the observations. The operation can be written as follows:

$$\text{Parent religiosity} +_d \text{self-parent conflict} \rightarrow \text{self religiosity},$$

where $+_d$ indicates the addition of deep structures, and $\rightarrow$ is a connecting operator separating the target and conforming observations. The analysis will bring the deep structure of self religiosity into conformity with the added deep structures of the parent religiosity and self–parent conflict observations.

The results, however, indicate that this operation was not very successful for the 80 individuals, yielding a PCC index of only 23.75% (19 of 80). The c value was quite low ($c = .09$, 1000 trials), but examination of the multigram in Figure 8.5 shows no clear pattern of association. It is also noteworthy that the PCC index decreased compared to the simpler model tested previously, parent religiosity $\rightarrow$ self religiosity. For the simpler model, the result was equal to 38.75% (31 of 80), and the c value was also more impressive ($<.001$). Unlike multiple regression, then, introducing more orderings of observations into an integrated model can hinder its accuracy. With multiple regression, R^2 is the most popular index of a model's efficacy, and it cannot decrease in value when variables are added to the model. In fact, it will likely increase by

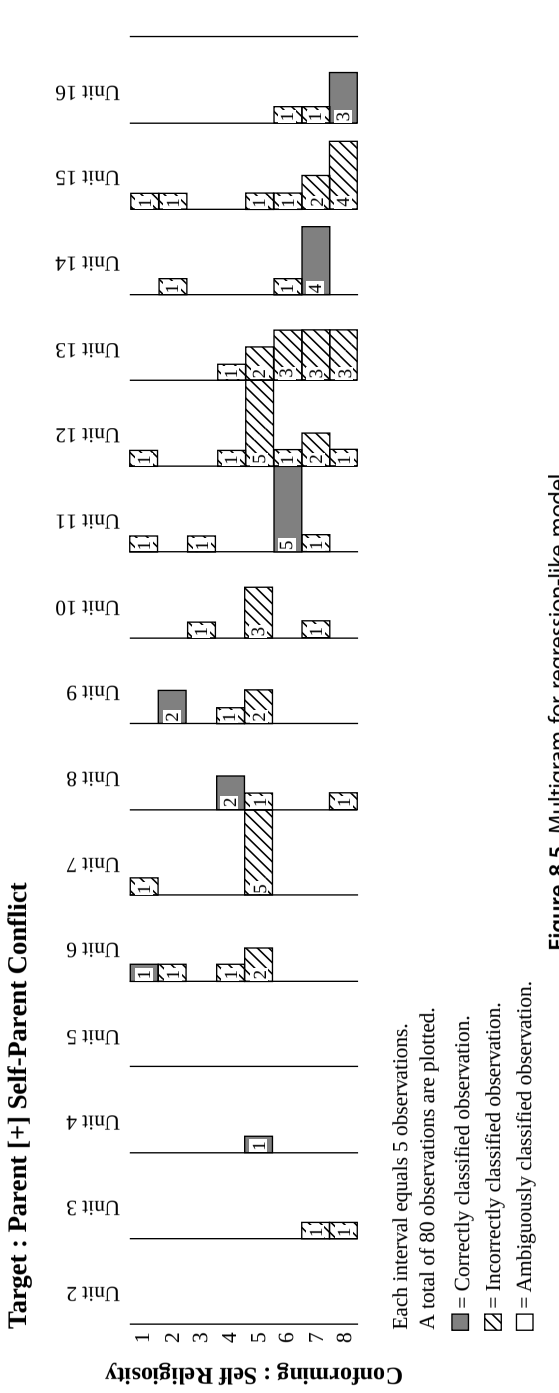

Figure 8.5 Multigram for regression-like model

a noticeable (although not necessarily statistically significant) amount when anything but random variables are added to the equation. Of course, in odd cases in which the number of variables is equal to the sample size minus 1 ($n - 1$), R^2 will by definition equal 1. None of these properties apply to the regression-like operations in observation oriented modeling.

Even more complex operations can be constructed and tested, but two considerations will likely limit their application. First, the number of columns in the multigram will grow large and may preclude a meaningful interpretation of the results, particularly if a paucity of observations are made. As noted previously, the analysis depends on sufficient observations being made for each of the units of observation. Second, as discussed in Chapter 9, deep structure rotation is not symmetric, and when the number of units in the target observations exceeds the number of units in the conforming observations, ambiguity is likely to result (see Chapter 2). Metaphorically, in such situations the model is attempting to "stretch" a small number of units of information into a larger number of units of information, and ambiguity often results. For the current example, no ambiguity was found even though the number of target units (15) was greater than the number of conforming units (8). If ambiguous classifications had been found, several or many of the bars in Figure 8.5 would not have been shaded or filled. Care must be taken, then, in constructing an integrated model with its requisite observations so that such asymmetries are clearly recognized and justified.

The previous example is sufficient, nonetheless, to demonstrate that regression-like operations can be constructed and evaluated in observation oriented modeling, but such operations are based on deep structure addition that does not assume continuous quantitative structure. It does imply that the units of observation possess an order that will be preserved when their deep structures are added. When the observations are continuous quantities, such as temperature and barometric pressure in Figure 8.1, traditional mathematical modeling is likely to prove more effective than observation oriented modeling, particularly if only one or two observations are available for each unit of measure.

MEASUREMENT ERROR

Any discussion of measurement is perhaps incomplete without an adjoining discussion of measurement error. Unfortunately, the previous

discussion must therefore in some sense be considered as incomplete because no analytical treatment of measurement error has been developed for observation oriented modeling. The two major models of measurement error, the classical true score model and generalizability theory, both rest squarely on the assumption of continuous quantitative structure.[14] Moreover, they are variance-based conceptualizations completely in sync with the Pearsonian–Fisherian tradition and are thus not generally suitable for observation oriented modeling. This fact can easily be seen in the standard presentation of reliability, r_{kk}, in the classical true score model,

$$r_{kk} = \frac{\sigma^2_{\text{true scores}}}{\sigma^2_{\text{true scores}} + \sigma^2_{\text{error}}} = \frac{\sigma^2_{\text{true scores}}}{\sigma^2_{\text{observed}}},$$

and in the variance components of a base model for generalizability theory,

$$\sigma^2_{\text{Xpi}} = \sigma^2_{\text{p}} + \sigma^2_{\text{i}} + \sigma^2_{\text{pi}} + \sigma^2_{\text{pi,e}}.$$

Fisher's analysis of variance is also used as the method for estimating the parameters in the latter model. The scientist who adopts observation oriented modeling must therefore, at least for the present time, be willing to give up a number of familiar practices, such as summing item responses in order to increase Cronbach's alpha or a G coefficient to some arbitrarily acceptable level (e.g., .70) or using the Spearman–Brown prophecy formula or results from a G study to estimate the number of items that must be added to a questionnaire to bolster its internal consistency. This is not to say that responses to a questionnaire, judgments of raters, or other observations cannot be combined in some manner but only that the formal psychometric treatments of averaging data cannot necessarily be transferred to observation oriented modeling.

A potentially positive effect of this change is the revelation that alternative methods must be pursued that focus more on making accurate judgments about observations than on fulfilling psychometric requirements

[14] The classical true score model dates back to the early 1900s and the work of Charles Spearman, but the classic text on the matter is considered by many to be *Statistical Theories of Mental Test Scores* by Lord and Novick (1968). Jum Nunnally's *Psychometric Theory* (1978) is also often cited for classical test theory. For generalizability theory, the text by Cronbach, Gleser, Nanda, and Rajaratnam (1972) is paramount. More recent and more readable treatments can be found in Brennan (2001) and Shavelson and Webb (1991).

that may be based on faulty assumptions.[15] Studies of short or single-item questionnaires provide one example of such alternative thinking. Matthias Burisch (1984), in a review article published in *American Psychologist*, showed that brief self-report personality questionnaires constructed without sophisticated psychometric analysis were just as valid as questionnaires with many more items. Regarding validity, Burisch focused primarily on the correlations between external criteria (e.g., peer ratings, grade point average, and psychiatric diagnosis) and scores on the brief and lengthy questionnaires. No meaningful differences were found when such correlations were examined; in other words, the brief questionnaires were shown to be just as effective as the lengthy questionnaires for predicting external criteria. Taking Burisch's findings further, an increasing number of studies are showing that single items can be just as predictive and just as stable over time as multiple-item questionnaires. Also, single items are not solely practical for personality psychologists; they have been shown to be effective in studies of job satisfaction, personal appraisals of pain and fatigue, and attitudes toward advertisements (Bergkvist & Rossiter, 2007; Butt et al., 2008; Nagy, 2002). Although based on aggregate statistical analysis, the guiding principle behind these studies is that accuracy (i.e., predictive validity) is the primary evaluative criterion for success, whereas internal consistency reliability (e.g., Cronbach's alpha), tied to traditional psychometric theories of error, is eschewed. Single items also permit the researcher to focus more clearly on a particular trait, attitude, concept, personal construct, etc. They may consequently be more easily incorporated into integrated models, much like the simple response formats described previously that are consistent with personal construct theory. The observation oriented modeler will therefore spend his or her time constructing a small number of items in the context of a well-reasoned model or devising ways to make precise observations rather than pursuing multiple-item questionnaires through modern psychometric theory.

Internal consistency can often be considered more simply and more profitably in terms of accuracy of agreement. For instance, a number of

[15] Perhaps the most stunning assumption underlying the classical true score model is the independence of test scores obtained from the same individual over repeated administrations. Lord and Novick (1968) did not completely shy away from this topic but, rather, spoke of "washing a person's brains" between test administrations (p. 29) as a way of conceptualizing what they referred to as the "weak" assumption of independence (p. 25). However, they ultimately relied on the purported usefulness of the classical true score model as their primary justification for accepting the assumption (p. 30).

people may be asked to rate the photographs of 10 target individuals regarding their perceived levels of narcissism, or a number of raters might be asked to score a set of responses to cards from the Thematic Apperception Test. Whereas modern psychologists would turn to a variable-based model and one of a number of available statistics, such as kappa or an intraclass correlation coefficient (ICC), to quantify the correlation between raters, Barrett (2009) has shown that a simpler route, entirely consistent with observation oriented modeling, can instead be adopted. With this approach, the goal is to assess the level of agreement between pairs of raters, provided the ratings possess the same format (i.e., the ratings are all on the same scale). No theory of measurement is invoked; rather, the original scales utilized by the raters are included in a series of simulations to judge their overall agreement in comparison to chance representations of responses to the same scales. The focus is therefore on the direct agreement of the raters, quantified in a standardized metric.

In observation oriented modeling, the degree of agreement between raters, or observations, is conveyed in the PCC index, which can be evaluated according to the *c* value. As an example, consider a study of 27 high school students who attended a Summer Science Academy sponsored by a university.[16] These students were judged by six individuals regarding their perceived likelihood of actually pursuing a career in science in the future. Specifically, each of the six judges rated the 27 students on a scale ranging from 1 to 10, with a rating of "1" reflecting certainty that the student would not pursue a career in science, and a rating of "10" indicating certainty that the student would pursue a career in science. The six raters supervised the students in various capacities during the course of the 2-week program and were therefore fairly well acquainted with the students. Following the Pearsonian–Fisherian tradition, the ICC for the six raters was found to be quite high (.89), thus indicating high agreement (maximum = 1). As demonstrated in Chapter 7, however, such standardized indices of effect size are often difficult to translate into practical terms because they are divorced from the original scales or the methods of collecting observations. With observation oriented modeling, the 10-point rating scale is respected as

[16] These data were collected as part of a Research Experience for Undergraduates (REU) opportunity at Oklahoma State University sponsored by the National Science Foundation. The undergraduate students participating in the REU program were studying the identities of high school students participating in a 2-week Summer Science Academy designed to pique their interests in pursuing careers in psychological science.

the method through which the researcher observed reality, and a simple matching analysis is instead conducted. This analysis tallies (counts) the number of exact matches between pairs of raters and converts these tallies to percentages. Using randomized versions of the deep structures of the ratings, a *c* value is also computed for each pair of raters. The results, as percentages, in Figure 8.6 show a high degree of variability in rater agreement. Only 3.70% of the ratings (1 of 27) for raters 1 and 2 matched, whereas 40.75% of the ratings (11 of 27) matched for raters 5 and 6. Overall,

Matching Analysis for ScienceRatings

Overall Results
Overall Percent Matches : 20.77
Overall c-value : 0.05

Number of Matches

	Rater1	Rater2	Rater3	Rater4	Rater5	Rater6
Rater1	27.00					
Rater2	9.00	27.00				
Rater3	8.00	7.00	27.00			
Rater4	6.00	10.00	6.00	26.00		
Rater5	3.00	3.00	5.00	3.00	27.00	
Rater6	1.00	4.00	3.00	4.00	11.00	27.00

Note. Imprecision = 0

Percent Matches

	Rater1	Rater2	Rater3	Rater4	Rater5	Rater6
Rater1	100.00					
Rater2	33.33	100.00				
Rater3	29.63	25.93	100.00			
Rater4	23.08	38.46	23.08	100.00		
Rater5	11.11	11.11	18.52	11.54	100.00	
Rater6	3.70	14.81	11.11	15.38	40.74	100.00

Note. Imprecision = 0

Overall c-values

	Rater1	Rater2	Rater3	Rater4	Rater5	Rater6
Rater1	.					
Rater2	0.00	.				
Rater3	0.01	0.02	.			
Rater4	0.02	0.00	0.02	.		
Rater5	0.50	0.50	0.13	0.34	.	
Rater6	0.94	0.27	0.50	0.17	0.00	.

Figure 8.6 Observation oriented modeling results for matching analysis of future scientist ratings

the percentages were low, and nearly half of the *c* values were not impressive ($\geq$.23), indicating that agreement between many pairs of raters was not much better than random pairings of their actual ratings. The overall percentage agreement averaged across all raters was only 20.77 even though the *c* value was impressively low (.05). Contrary to the high ICC result (.89), the agreement between raters was therefore not impressive when considering the actual values on the 10-point scale.

Traditional psychometric approaches assume that the ratings are structured as continuous quantities. With observation oriented modeling, they can be considered to represent the raters' ordinal judgments of greater or lesser likelihood of pursuing a scientific career. With this in mind, the poor agreement between raters can be further explored in two different ways. First, the observations can be regrouped into smaller numbers of units, much as was done previously with the religiosity items. Perhaps if the scales were divided into equal halves (1–5 and 6–10), the agreement would be higher. Such changes to the scale or method of observation would of course best be guided by an integrated model. Second, the matching analysis can be conducted with varying degrees of unit imprecision. For instance, a match for any two raters could be considered to occur when the discrepancy between their ratings is less than or equal to 2 deep structure units that are assumed to be ordinally arranged. In other words, perhaps the raters were accurate within ± 2 deep structure units. Following this strategy for the current six raters and using a unit imprecision value equal to ± 2, the results shown in Figure 8.7 reveal dramatic improvement. The minimum agreement is 44.44%, whereas the maximum agreement is 88.89%, with almost all of the *c* values less than .10. The overall agreement is 65.77%, with a *c* value of .03. The notion of precision, construed as applying to ordered categories, can therefore be incorporated in the analysis, and by staying close to the observations as they are recorded on the 10-point scales, a much more meaningful interpretation of the ratings emerges. The overall agreement between raters was fairly high when their ordinal judgments were not considered as perfectly precise. Moreover, some pairs of raters showed more agreement than others, and these differences would be worthy of further exploration, particularly as they might relate to who among the 27 students actually pursue a career in science in the future. Such thinking works against any inclination to add the ratings together into some sort of composite because such an operation would lose a great deal of information in the observations; however, this is often the purpose behind the ICC—to provide a rationale for summing observations. In observation oriented modeling, unless an integrated model is available that expressly incorporates a sum score, a richer

Matching Analysis for ScienceRatings

Overall Results
Overall Percent Matches : 65.77
Overall c-value : 0.02

Number of Matches

	Rater1	Rater2	Rater3	Rater4	Rater5	Rater6
Rater1	27.00					
Rater2	19.00	27.00				
Rater3	24.00	18.00	27.00			
Rater4	17.00	18.00	21.00	26.00		
Rater5	19.00	12.00	19.00	16.00	27.00	
Rater6	15.00	12.00	13.00	16.00	24.00	27.00

Note. Imprecision = 2

Percent Matches

	Rater1	Rater2	Rater3	Rater4	Rater5	Rater6
Rater1	100.00					
Rater2	70.37	100.00				
Rater3	88.89	66.67	100.00			
Rater4	65.38	69.23	80.77	100.00		
Rater5	70.37	44.44	70.37	61.54	100.00	
Rater6	55.56	44.44	48.15	61.54	88.89	100.00

Note. Imprecision = 2

Overall c-values

	Rater1	Rater2	Rater3	Rater4	Rater5	Rater6
Rater1	.					
Rater2	0.01	.				
Rater3	0.00	0.02	.			
Rater4	0.03	0.01	0.00	.		
Rater5	0.01	0.51	0.01	0.06	.	
Rater6	0.17	0.51	0.38	0.06	0.00	.

Figure 8.7 Observation oriented modeling results for matching analysis with two units of imprecision

understanding of the observations is preserved by remaining faithful to the scale or method used to obtain the observations and analyzing the observations in a way that is relatively free of assumptions.

LATENT VARIABLES

Combining responses for primarily psychometric reasons (e.g., increasing internal consistency) also raises questions about latent variable models commonly employed in the social sciences. It was previously stated that

Aristotle and Aquinas understood the concept of unity to form the basis of all measurement. In psychology, items on questionnaires or judgments obtained from multiple raters are typically combined because they are believed to express a single ability, trait, or some other attribute, such as intelligence, extraversion, or depression. The items are, in a sense, considered as "one." Often, one of two latent variable representations is used to express this unity: the common factor model (shown in Figure 8.8) or the item response theory (IRT) model. Despite their popularity, however, these models have several features or properties that make them incompatible with observation oriented modeling. First and foremost, they are variable-based models built on the assumption that latent variables are structured as continuous quantities satisfying Hölder's axioms. This assertion is true even for IRT, which is used to model dichotomous or polychomotous observations. The continuing failure of psychometricians and psychologists to acknowledge this key assumption of IRT and come to grips with its negative impact on theory development is what prompted Michell (2008) to ask, "Is psychometrics a pathological science?"

The common factor model, which stands on its own and forms the basis of structural equation modeling, is also known to be indeterminate in nature. This property is best known among psychometricians as factor score indeterminacy (Grice, 2001).[17] Practically, the consequence of this property is that scientists using the common factor model have no way to unambiguously connect the scores on a set of variables (the squares in Figure 8.8) to the presumed, underlying latent variables (the ellipse in Figure 8.8). Depending on the features of a given data set, different degrees of indeterminacy may be present. For instance, the Freedom from Distractibility factor (a latent variable) from an earlier version of the Weschler Intelligence Scale for Children, the WISC-III, was shown to have a very high degree of indeterminacy (Grice, Krohn, & Logerquist, 1999). This result meant that two hypothetical researchers could devise ways of ordering a group of children on the latent variable that would be entirely different. Whereas one researcher might score a particular child, for instance, as highly intelligent, the second researcher might score the same child as possessing only average intelligence. Because of the high degree of indeterminacy, both scores

[17] An extensive debate and discussion regarding factor scores was published in 1996 in *Multivariate Behavioral Research*. The lead article was Maraun (1996a). The discussion articles were published in the same issue. A comprehensive list of citations (up to 2001) relevant to factor scores and factor score indeterminacy can be retrieved from http://psychology.okstate.edu/faculty/jgrice/factorscores.

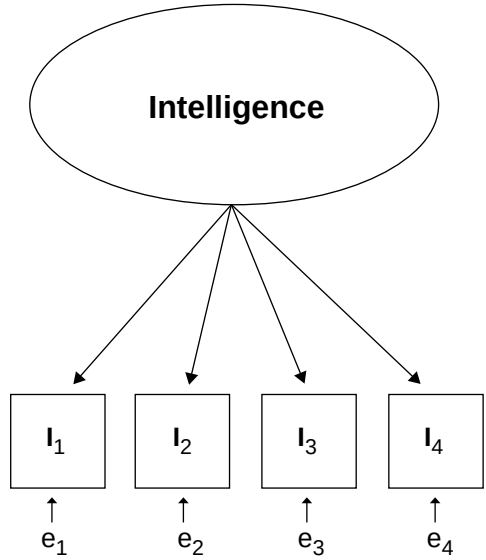

Figure 8.8 Example common factor model

would be entirely consistent with the latent variable. Despite this stunning fact, factor score indeterminacy was ignored by the developers of the WISC-III, and it has similarly been ignored by virtually all psychologists for the past 80 years.[18] Indeterminacy was shown to be inherent in the mathematics of the common factor model in the 1920s. Stated simply, indeterminacy is a matter of solving a set of equations in which the number of unknowns exceeds the number of equations. In such situations, an infinite number of solutions, all equally valid, may potentially be found, which helps explain why a child can be validly scored as possessing both high and only

[18] With a chance to perhaps turn things around and bring factor score indeterminacy to the attention of psychologists, the scholars responsible for holding a landmark conference celebrating the 100th anniversary of Spearman's factor analytic model failed to invite a single author to present on the topic. Factor score indeterminacy was also hardly mentioned in the edited book resulting from the conference (Cudeck & MacCallum, 2007). When such fundamental issues are ignored by leading quantitative social scientists, it appears that most are happy to repeat the errors of the past: "At the very least, are we going to insist that reputable textbooks have clear discussions of factor indeterminacy, and that computer programs provide determinacy indices for 'factors'? Or are we going to repeat the errors of the past?" (Steiger, 1996, p. 629). On the other hand, Kenneth Bollen (2002) acknowledged the indeterminacy of latent variables, although he erroneously speaks of "estimating" latent variable scores and falls short of recommending the computation of indices that quantify the degree of indeterminacy in a particular set of observations and model.

average intelligence for the same common factor. The fact that factor score indeterminacy has been ignored does not erase the fact that indeterminacy is part and parcel of the common factor model, as well as structural equation modeling and other latent variable models, and therefore the ambiguity of connecting observations to latent variables remains (Bollen, 2002).[19]

A third, more eclectic issue is the widespread misunderstanding of the nature of latent variables. Much of the confusion stems from referring to latent variables as "unobservable," "unmeasurable," or "error free" or from replacing the word "variable" with other words, thus yielding so-called "latent traits" or "latent constructs," which are then taken to be latent variables. The line of thinking followed by many who utilize latent variables begins with an attribute, such as intelligence, that is considered "latent" because it is never observed directly—only its effects are observed, such as responses to a set of math problems. The latent attribute is then considered to correspond to a latent variable that unifies the observations, not just pragmatically but causally as well (Borsboom, Mellenbergh, & Van Heerden, 2003). As Michael Maraun and Peter Halpin (2008) have noted, this line of reasoning usually results in confusing things in nature, which can or cannot be directly observed, with variables, which are generically defined as conceptual, numeric placeholders.[20] The distinction can be clarified by considering a basketball and a bowling ball. A person completely unfamiliar with each ball may visually inspect the two and imagine picking up the bowling ball with greater ease because of its smaller size. Upon picking up the balls, however, to the person's surprise the bowling ball is extremely heavy compared to the basketball. The person may continue to examine the balls, now in a tactile manner, and may even attempt to open them using different tools for further inspection. At any point in this process, the person does not observe "heaviness" in the same way that each ball's color, odor, material composition, etc. are observed. The quality heaviness can thus be considered as more or less observable, but a variable has not yet been posited. If the person next devises a method for analogously measuring heaviness using a spring-loaded scale and then states, "Let x equal the weight (i.e., measured heaviness) of the object as read from the scale," at that point the person has defined a variable, which is fundamentally different than

19 Bollen (2002) lists indeterminacy as a property of the latent variable models he discusses in his article.

20 For a fuller treatment of Maraun's philosophical views regarding social science, see Maraun, Slaney, and Gabriel (2009).

considering heaviness as a quality, for instance, related to an object's material composition, density, or mass. Similarly, a psychologist may posit intelligence as a quality or attribute that may be more or less directly observable, but once it is considered as a latent variable, like the one shown in Figure 8.8, then it must be regarded as a variable per se, which makes it a simple linear function of other variables (the squares in Figure 8.8). A latent variable can no more be predicated as "unobservable" or "unmeasurable" than numbers read from a spring-loaded scale designed to measure heaviness.[21] In other words, there are no per se latent variables. Moreover, to describe latent variables as "error free" is to confound the task of taking measurements, which may vary in accuracy both randomly and systematically, with the mathematical operations involved in computing a latent variable. This is most clearly seen in the common factor model in which the factors (latent variables) are computed from the observed variables as weighted composite scores that are—much to the chagrin of latent variable advocates—conceptually equivalent to the allegedly error-laden components in a principal components analysis.[22]

This picture can be made even fuzzier if latent variables are considered as random variables. The reason for the increased confusion is that a random variable is not itself a variable per se; rather, it is a function connecting a domain to a range. It is further even debatable if a latent variable, considered as a random variable, is actually a function at all because it fails to uniquely map the elements of a domain (item scores) to elements in the corresponding range (factor scores). As stated by Peter Schönemann (1996a),

> *Now, while the relation from the sample space to the test space is many–one, and thus a map [function], their relation from the test space to the factor space is, in view of the indeterminacy, many–many and thus not a map. Hence, the composite relation from the sample space to the factor space is not a map either, and factors cannot be random variables by the conventional definition.*
>
> ***(p. 574)***[23]

21 This subtle confusion has been part and parcel of the classical true score model as well since at least the time of Lord and Novick's (1968) classic text: "Since observed scores are directly observable and true scores and error scores are not, we shall express the various parameters of true and error scores in terms of parameters of observed scores" (p. 55). Again, it is difficult to imagine how a score, which is a placeholder, can be profitably construed as "unobservable."

22 See Mauran (1996b). The common factor model is compared to principal components particularly on pages 676–678.

23 See also Schönemann (1996b).

It seems that a latent variable might best be defined as a relation rather than as a function (i.e., random variable).

The conclusion to be drawn is that although a number of psychometricians and methodologists in psychology have been pushing for the universal application of latent variable modeling, the untested assumption of continuous quantitative structure, the methodological and analytical problems (i.e., indeterminacy), and the deep conceptual confusion regarding latent variables suggest it is something of an intellectual cul-de-sac.[24] A different route such as observation oriented modeling is therefore warranted. Philosophically, positing attributes, forces, objects, etc. that are not readily observable is an important aspect of discovery in natural science. Once posited, however, these hypothetical phenomena must find their value in formal models, testable hypotheses, and attempts at their measurement through either the analogous transfer of dimensive quantities (as described previously) or careful observation and counting. In this way, the intellectual move is always from what is lesser known to what is better known about the things of nature, but to paraphrase Maritain (see Chapter 6), the move is always in the realm of sensible operations. Positing latent variables that cannot be unambiguously connected to the observations from which they are constructed is clearly a move in the opposite, and wrong, direction, which explains why so much of latent variable modeling revolves around endless arguments of choosing the appropriate index of model fit.[25] In many ways, observation oriented modeling represents a simplification in the sense that unwarranted assumptions and overly abstract concepts are suspended unless they demonstrably yield impressive improvements to the accuracy of an integrated model's predictions. Without continuous quantities and a formal model of measurement error (based on unrealistic assumptions) to rely on, the psychologist must return to the basic principles of scientific investigation: removing error through carefully controlled observations and the repetitious gathering of observations that are considered suspect. Recall Mendel's repeated observations of a strain of crossed peas he suspected was tainted (see Chapter 5). Methodologically, the move is away from the Pearsonian–Fisherian tools of analysis and null hypothesis significance testing to techniques that provide clear tests of

24 Two examples favoring latent variables are Borsboom (2008) and Pearl (2000).

25 The published papers that could be cited with regard to fit indices are legion. The reader is encouraged to peruse the journal *Structural Equation Modeling* or to start with the paper by Marsh et al. (2004).

integrated models. Observations that are not demonstrated to possess continuous quantitative structure are considered as countable discrete units or ordinal judgments, thus suggesting that the world of finite mathematics may provide the most profitable tools of modeling and analysis. Such methods are being used by modern psychologists, but their potential may not be fully realized until they are completely divorced from the Pearsonian–Fisherian framework and instead wedded to philosophical realism and tenets similar to those found in observation oriented modeling.

This page intentionally left blank

CHAPTER 9

Cause and Effect

Contents

Introduction 183
Causality and Causation 184
The Randomized Controlled Experiment 187
Mediation Models 190
Person-Centered Mediation Analysis 195
Logical Hypothesis Testing 201
Final, Formal, and Efficient Causes 206
Transition to the Four Causes 214

INTRODUCTION

In a comprehensive review article discussing the status of research on the causal mechanisms underlying psychotherapy, Alan Kazdin (2007) observed,

> *There has been enormous progress in psychotherapy research. This has culminated in recognition of several treatments that have strong evidence in their behalf. Even so, after decades of psychotherapy research, we cannot provide an evidence-based explanation for how or why even our most well-studied interventions produce change, that is, the mechanism(s) through which treatments operate.*
>
> ***(p. 1)***

In the context of the arguments presented in this book, Kazdin's statement is a clarion call for the development of integrated models in psychology. His statement also casts a bright light on the centrality of causality in scientific research. Demonstrating that one type of therapy is superior to another with regard to effectiveness is not sufficient; rather, the goal of psychotherapy research should be to understand the causal mechanisms underlying the different types of therapy. After decades of research, why haven't such causal models of therapeutic change been developed and tested?

In many ways, an answer to this question has already been given in previous chapters. The abstract nature of null hypothesis significance testing, the practical and theoretical inutility of effect sizes, and the failure of psychologists to confront fundamental principles of measurement

Observation Oriented Modeling
ISBN 978-0-12-385194-9, Doi:10.1016/B978-0-12-385194-9.10009-X
© 2011 Elsevier Inc.
All rights reserved.

have all been shown to prevent the proper development of integrated models in psychology. Because integrated models are essentially causal models, their scarcity suggests that causality is not central to the ways psychologists construct and test their theories. No doubt innumerable published papers and research textbooks utilize cause-and-effect language, but such manifestations of causal reasoning are steeped in the Pearsonian–Fisherian tradition and are therefore limited and potentially obfuscating. Indeed, in this chapter the predominant view of causality held by most psychologists, which has its origins in Cartesian idealism, is discussed and contrasted with the view held by Aristotle and Aquinas and adopted in observation oriented modeling. The randomized controlled experiment and mediation analysis are then discussed, and alternative methods of analysis are introduced. Logical hypothesis testing, which is a unique analytical tool offered in observation oriented modeling, is then demonstrated. Finally, Aristotle's efficient, formal, and final causes are discussed in the context of several examples. As with previous chapters, the goal is to elucidate the limitations of the Pearsonian–Fisherian tradition while highlighting the strengths of observation oriented modeling and the philosophical realism upon which it rests.

CAUSALITY AND CAUSATION

For Aristotle, causality was central to science, which he considered to be demonstrable knowledge (*epistēmē* in Greek, translated to *scientia* in Latin) of nature through its causes. According to William Wallace (1996, p. 292), by "demonstrable" Aristotle meant logical demonstration via categorical syllogism. Consider the following example from Wallace's book *The Modeling of Nature*, regarding one of the original arguments for the sphericity of the earth:

S		**M**		**P**
The earth	is	a body that casts a circular shadow on the moon during an eclipse	is	a sphere

The syllogism is composed of a subject (S), middle term (M), and predicate (P) and is read in the mode *Barbara*; specifically, "A body that casts a circular shadow on the moon during an eclipse is a sphere; the earth is a body that casts a circular shadow on the moon during an eclipse; therefore, the earth is

a sphere" (p. 283).[1] In this particular example, the predicate carries the only cause in the argument, and it is a formal cause according to Aristotle's classification because it refers to a geometric form (i.e., a sphere). If the content of the middle and predicate terms is true, and the linkages between terms are true, then the conclusion immediately and necessarily follows. In this sense, the knowledge is certain, and here it demonstrates an important aspect of the formal cause of the earth.

The modern psychologist or social scientist may cringe at the suggestion of certain knowledge of nature, and of course the truth value of the terms may be debated, particularly in the context of an ancient understanding of the universe. Moreover, various suppositions come into play. Some are mundane; for example, it is supposed that persons judging this demonstration understand what a sphere is and that under normal circumstances it will cast a circular shadow. More sophisticated suppositions are that at great distances light travels uniformly and in straight enough lines so as to create a shadow that corresponds to a spherical earth. Of course, the truth of the middle and predicate terms may also be established via further reasoned argument or careful observation and controlled experimentation. Most historical accounts point to the Middle Ages and Europe as the cradle of modern science, with its developing appreciation of controlled methods of experimentation and formal (often mathematical) modeling. According to Wallace, however, what has been lost in the historical analysis is the importance of logical demonstration with regard to the acquisition of scientific knowledge. Following this loss has been the ascendancy of the view that science is essentially probabilistic in nature and incapable of yielding the *scientia naturalis* or *epistēmē* envisioned by Aristotle. The terms in the previous syllogism, for instance, are more commonly framed in the logic of propositions:

> *If the earth is a sphere, then the earth will cast a circular shadow on the moon during a lunar eclipse.*
>
> *The earth casts a circular shadow on the moon during a lunar eclipse.*
>
> *Therefore, the earth is* probably *a sphere.*

The word "probably" is included in the conclusion because other explanations for the circular shadow are not logically ruled out. In other words, the spherical shape is sufficient for the circular shadow, but it is not held to be necessary. If such uncertainty in propositional logic is construed as the

[1] Wallace provides numerous additional examples of demonstrable knowledge from different sciences in his *From a Realist Point of View* (1983).

essence of science, then the importance of causality is undercut because causes themselves are defined as necessary for effects. Rephrasing the proposition "If and only if the earth is a sphere, then a circular shadow will be observed" would reflect such necessity and be equivalent to the previous categorical syllogism, thus keeping causality at the center of reasoning. As stated by Wallace (1996),

> *That is to say, if and only if the earth is a sphere will a proper causal explanation be given of the various observational phenomena.... But stating the arguments in the logic of propositions rather than in the logic of terms unfortunately has the effect of blocking out such causal connections.*
>
> ***(p. 285)***

This is not to say that propositional logic is not a valuable tool in scientific discovery but to simply note that it is only a tool and not the defining feature of science. Karl Popper's falsification criterion of course rests on such propositional logic and, along with other purely deductive perspectives (e.g., the hypothetico-deductive model), must therefore be seen as failing to provide an adequate definition of science (pp. 244–249).[2]

The point being made here is that to the extent they view science exclusively through the lens of propositional logic, psychologists will hold fallibility and probability above causality and modeling. Such a view is disadvantageous because it works directly against the development of integrated models that elucidate the structures and processes of nature and of human action.[3] However, such models provide the very causal explanations called for by Kazdin at the beginning of this chapter. Why do particular therapies produce change while others do not? Why do some people benefit from therapy while others do not? Scientific answers to such questions of "why" can be provided only insofar as general principles of causality are invoked, and modeling provides the necessary

[2] Rom Harré discusses what he calls the "myth of deductivism" in the first chapter of his *Principles of Scientific Thinking* (1970). Peter Manicas challenges the deductive–nomological (i.e., the covering law) model of explanation in his *A Realist Philosophy of Social Science* (2006). Finally, in describing the "demonstrative regress" for acquiring scientific knowledge, Wallace (1996, pp. 300–308) includes both inductive and deductive steps.

[3] "Characteristically, structure is presented diagrammatically (by pictures and models), and the possibilities of change sententially, as conditional statements. By considering only conditional statements, and forgetting the structural 'picture' from which they have been abstracted, one gets the characteristic 'event' view of the world since it is just the successive states of things, but not the things themselves, that conditions describe" (Harré, 1970, p. 13). The "event view" referred to by Harré is very much like what has been described as variable-based thinking in this book.

heuristic for specifying the material, efficient, formal, and final causes of nature. Fallibility and probability are concepts that offer no such heuristic and instead shore up the Pearsonian–Fisherian approach toward building and testing variable-based models using null hypothesis significance testing and other inferential statistical methods. Gerd Gigerenzer (2004) criticized what he referred to as the ritualistic application of statistical methods in the social sciences. Unfortunately, he did not go far enough in his critique when announcing that if the rituals can be stopped, social science might "enter the dawn of statistical reasoning" (p. 604). From the perspective of observation oriented modeling, psychologists and other social scientists should look forward to the twilight of statistical reasoning and a return to the commonsense realism of the early scientists. Causality can then take center stage in efforts to model nature from the standpoint of Aristotle's four causes (Wallace, 1983).[4] Psychologists might then begin to recognize the limitations of viewing causation only through the lens of the randomized controlled experiment.

THE RANDOMIZED CONTROLLED EXPERIMENT

For many psychologists, and scientists from other disciplines as well (e.g., medicine and agriculture), the randomized controlled experiment is considered the "gold standard" of evidence in supporting causal claims. In a basic randomized controlled experiment, two discrete groups are determined—the control and experimental groups. The latter group receives a treatment of some sort, such as a new drug, whereas the control group receives no treatment, a placebo, or a competing treatment depending on the goals of the scientist and previously conducted research. Individuals are randomly assigned to each of the two groups (the independent variable) and compared on a relevant outcome measure (the dependent variable) delivered after the treatment phase of the study.

Randomized controlled experiments are commonplace in psychology and the social sciences, and they can provide evidence for causality when understood and used properly. Far too often, however, the randomized experiment is treated haphazardly, leading researchers into various confusions regarding the method's capabilities and limitations. The most common

[4] In Essay VI, Wallace (1983) provides numerous examples of demonstration from various sciences that incorporate all four of Aristotle's causes. See also Chapters 8 and 9 of his *The Modeling of Nature* (1996).

misunderstanding is that random assignment controls for the effects of individual difference variables in a single study.[5] Suppose researchers studying the efficacy of a new drug assigned only Caucasians to the experimental group and only Hispanics to the control group. Using variable-based language, ethnicity is an individual difference variable, and in this contrived scenario it is a confounding variable. The people in the two groups (drug vs. placebo) now differ systematically in terms of their ethnicity, which can be causally implicated in explaining differences in the outcome. For instance, Caucasian individuals may have a different metabolic reaction to the drug compared to Hispanics. The success of the drug as demonstrated in the study therefore depends on this ethnicity variable, and with such a confound there simply is no way to make an unambiguous causal claim.

By randomly assigning individuals to two groups, researchers hope to control for such individual difference variables—using variable-based language of the Pearsonian–Fisherian approach. The number of individual difference variables imaginable is also great, thus making random assignment a potentially powerful method of controlling confounds. Differences in age, ethnicity, social economic status, height, weight, eye color, personality traits, intelligence, etc. can all ostensibly be controlled by simply randomly assigning individuals to the groups that constitute the independent variable. Many social scientists fail to remember, however, that equality between groups of randomly assigned individuals is only ensured *in the long run*. For any single study, random assignment may fail completely, resulting in large differences between groups on a number of individual difference variables. The researchers discussed previously, for instance, could use a simple coin toss or random number generator to assign participants to the drug and food groups but still assign disproportionate numbers of Caucasians and Hispanics to the two groups. The principle is very simple and understandable through the common experience of tossing, as randomly as possible, a coin. If the

[5] Further discussion of the limitations of random assignment can be found in Cook and Campbell (1979). This book is considered a classic among social scientists and provides discussions of various threats to conclusions drawn from randomized controlled experiments and nonexperimental studies. The purported philosophy underlying Cook and Campbell's arguments is referred to as an evolutionary critical realistic perspective. However, their overreliance on probability, J. S. Mill's theories, and statements such as "Our concern is less with understanding the causal determinants of a phenomenon than in assessing how reasonable it is to assume that particular manipulable causes usually produce a given result" (p.15) suggests an idealistic foundation that is incompatible with observation oriented modeling.

coin is tossed 20 times, the single most likely outcome is 10 heads and 10 tails, but no one would be surprised if 13 heads and 7 tails were observed. Such deviations from the expected 50/50 split between heads and tails are a natural part of the sampling process. As the number of tosses increases, or as the number of 20-trial experiments increases, however, large deviations from an even split between the proportions of heads and tails would lead to suspicion regarding the randomness of the tossing procedure. As simple as this principle may seem, it is easy to forget, even for seasoned academics, as shown by the following statement by Philip Zimbardo describing his famous prison experiment:

> *Ultimately, we were left with a sample of 24 college students from the U.S. and Canada who happened to be in the Stanford area and wanted to earn $15/day by participating in a study. On all dimensions that we were able to test or observe, they reacted normally. Our study of prison life began, then, with an average group of healthy, intelligent, middle-class males. These boys were arbitrarily divided into two groups by a flip of the coin. Half were randomly assigned to be guards, the other to be prisoners. It is important to remember that at the beginning of our experiment there were no differences between boys assigned to be a prisoner and boys assigned to be a guard.*
>
> ***(Retrieved August 2008 from http://www.prisonexp.org)***[6]

Perhaps Zimbardo meant to state the boys' scores on the various tests were relatively homogeneous, but taken at face value, the statement promises perfect equality (i.e., "no differences") between the prisoners and guards on the basis of random assignment. Such a promise will normally be broken in a single study and can in fact only be fulfilled if realized in the long run.

An important consequence of this property of random assignment is that exact replication must again come to the foreground. Without exact replication, even a study as intriguing and sensational as the Stanford prison experiment is severely limited in its scientific value, which is to say that it does little to further our understanding of the causal forces at work in human behavior.[7] Casting the experiment even in the simplest metaphorical terms, it cannot be said that a "bad apple" did not spoil the "barrel" of guards, thus resulting in their abusive behavior, even though Zimbardo's study is often interpreted as showcasing the importance of situational over personality factors in causing harmful behavior.

[6] See also http://www.bbcprisonstudy.org/index.php.

[7] The Stanford prison experiment suffered from numerous additional problems that severely damaged its scientific utility; most notably, Zimbardo participated in the study in a supervising role. See Carnahan and McFarland (2007).

The randomized controlled experiment can be traced back to the work of John Stuart Mill in the 19th century. This history is significant because, like David Hume, Mill held an idealistic and skeptical view of causality. It is idealistic insofar as the world of sense impressions is the starting point for discussing how people come to know the things of nature as well as their interrelationships. Because the domain of sense impressions, which is ordered according to space and time, is the starting point, a person can only be certain of the logical and temporal co-occurrences of events in nature as they are present in the mind. It is in this sense that Mill's view is skeptical and consistent with Hume's efforts to replace the notion of causality with that of causation. The difference between the two concepts is that the former views causes as inhering in the things of nature, whereas the latter views causes as consistent associations of sensory data present only in the minds of persons. This is again reminiscent of Pearson's "shadow table" (see Chapter 5), thus pointing to a common set of philosophical presuppositions underlying modern statistics and research practices. In any case, Mill's famous inductive methods for establishing the causal connection between sense impressions fits snugly with the Pearsonian–Fisherian framework. His method of concomitant variation is often paired with Pearson's correlation and chi-square statistic, whereas his joint method of agreement and difference is invoked when treatment and control groups are compared on a dependent variable with Fisher's analysis of variance. Therefore, like Stevens' four scales of measurement (see Chapter 8), Mill's methods have served as effective supports for the Pearsonian–Fisherian approach and variable-based modeling. This is not to say that Mill's methods are of no value; indeed, they can be useful if properly understood and used for highly specific purposes, such as determining the aggregate-level efficacy of a treatment or identifying necessary preconditions for particular effects, as will be done later with logical hypothesis testing. It is instead to argue that the philosophical idealism that permeates these methods has led social scientists astray by restricting their view of causality to efficient causes existing only as temporally associated sense impressions in the mind.

MEDIATION MODELS

In the late 1900s, the popularity of mediation variable-based models ascended rapidly in the social sciences.[8] A mediation model is one that

[8] The paper by Baron and Kenny (1986) spurred interest in mediation models. David MacKinnon (2008) offers a thorough introduction to these types of variable-based models.

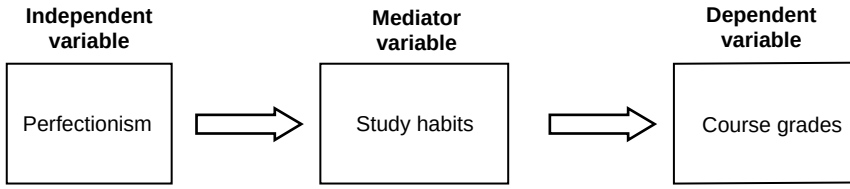

Figure 9.1 Variable-based mediation model

temporally connects three variables in an ordered sequence, such as the model shown in Figure 9.1. Because of their dependence on time, and the assumption that each variable prior in the sequence is a necessary initiator of the posteriorly connected variable, these models normally address efficient causes. Unfortunately, a review of published studies incorporating mediation models found that most researchers fail to adequately address the time-dependent nature of the alleged causes, which suggests that many tests of variable-based mediation may be little more than exercises in scripted statistical analyses that actually obfuscate rather than elucidate causal mechanisms (Gelfand, Mensinger, & Tenhave, 2009).[9]

Variable-based mediation models may also hinder proper causal reasoning by masking the logical disconnect between the level at which the models are tested and the level at which the causal forces are known to operate. This issue is similar to that raised in Chapter 7 regarding effect sizes. Recall that effect sizes based on between-person aggregate statistics typically do not represent the effects observed at the level of the observations—that is, at the level of the individuals in the study. Similarly, the level at which causes are normally considered to operate for psychologists is the level of the person, whether the person is alone completing a task or interacting with others. Causal processes are not normally posited to operate at the level of a computed aggregate statistic, such as a mean, variance, covariance, or correlation; however, that is the level at which mediation models are typically tested. In fact, from a realist perspective, causes cannot operate at such an abstract level because they must instead be recognized as inhering in the things (i.e., persons, objects, and things) of nature. This idea is not new: William Stern correctly argued in 1900 that psychology is indeed a discipline of persons, not of aggregates and variables (Lamiell, 2003, 2009).

A few methodologists have understood the limitations of variable-based mediation models and developed alternative person-centered approaches.

[9] Gelfand et al. (2009) provide an excellent summary of the various pitfalls associated with mediation models.

These approaches are more consistent with the philosophy and methodology of observation oriented modeling and are discussed later. It is first necessary, however, to demonstrate an alternative approach toward analyzing traditional mediation models by taking an example from David MacKinnon's book *Introduction to Statistical Mediation Analysis* (2008). MacKinnon reports a mediation model in which the causal connection between the quantity of consumed grilled meat and the incidence of pancreatic cancer is examined. The mediating variable in the example is the amount of fatty acids in a person's blood. Although the data are contrived, they nonetheless offer an interesting example to which observation oriented modeling can be applied. The model can be seen in variable-based form in Figure 9.2. The range of scores for the consumed grilled meat is 1 to 7 (the number of days for which grilled meat is typically consumed). The amount of fatty acids in the blood is recorded on a scale of 1 to 7 (1 = low, 7 = high), and the incidence of pancreatic cancer is binary (yes or no). With observation oriented modeling, an effect must conform to a cause, and the analysis begins with the last two sets of ordered observations in the sequence: fatty acid level (efficient cause) and incidence of cancer (effect). In the traditional Pearsonian–Fisherian approach, the two variables could simply be standardized and their correlation determined, or a logistic regression model could be used to estimate the relationship between the two variables in some hypothetical population. In observation oriented modeling, the goal is to conform the cancer observations, ordered in two units, to the fatty acid observations, ordered in seven units. The result of this analysis from the Observation Oriented Modeling software is shown in the multigram in Figure 9.3. Given the asymmetry involved with conforming two units to seven, a large number of observations were unsurprisingly classified ambiguously (52%) or incorrectly (47%).

The analysis and the lens through which the observations are being viewed now raises an important question: How can a countable quantity be causally translated (as an efficient cause) into a binary effect? Without an integrated model, a definitive answer cannot be given, but it might

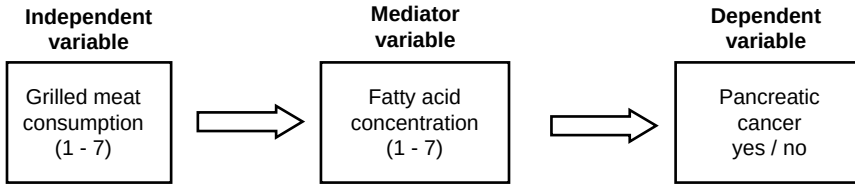

Figure 9.2 Mediation model for meat consumption and pancreatic cancer

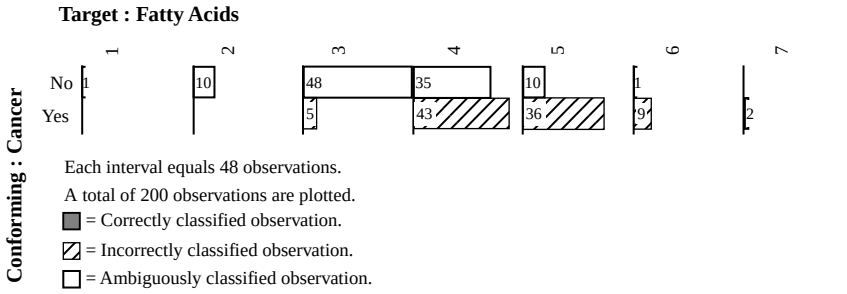

Figure 9.3 Multigram conforming cancer (two units) to fatty acids (seven units)

nonetheless be reasonable to posit some type of threshold that a person must cross when consuming grilled meat in order to develop pancreatic cancer. Examination of the multigram in Figure 9.3 suggests a threshold of four servings per week. In other words, if a person consumes four or more servings of grilled meat per week, then he or she will develop pancreatic cancer, or the incidence will at least be greater given that other important observations may be missing. The fatty acid observations are thus regrouped into two categories: 1–3 servings per week and 4–7 servings per week. The pancreatic cancer observations are then rotated to conformity with these grouped observations. The results in the multigram in Figure 9.4 show a high number of correctly classified observations (percent correct classification (PCC) = 74.50; c value $< .001$, 1000 trials; classification strength index (CSI) values between .91 and 1) and support the notion of a threshold that must be crossed before the fatty acids will initiate (efficient cause) cancerous growth in the pancreas. Of course, not all of the observations

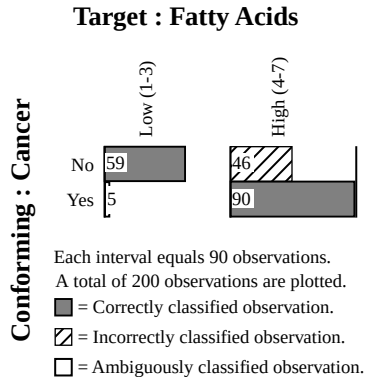

Figure 9.4 Multigram conforming cancer (two units) to fatty acids (two units)

were correctly classified because 46 persons were observed to have high amounts of fatty acids in their blood but did not develop the cancer, and 5 individuals were observed to have low amounts and still developed the cancer. Future research guided by an integrated model must therefore support this potential causal connection. Such research might involve simply ensuring the accuracy of the observed fatty acid levels or take on the more difficult task of searching for additional causes that work to offset or enhance the causal influence of the fatty acids.

Working back to the second link in the causal model, the 149 observations that were correctly classified in the previous analysis are next brought into conformity with the initial cause (i.e., grilled meat consumption). Because the fatty acid observations have now been regrouped, the analysis again involves rotating two-unit ordered observations into conformity with seven-unit ordered observations. The resulting multigram (not shown) revealed numerous ambiguously classified observations and supported the same reasoning presented previously regarding the application of a threshold to the meat consumption observations. These observations were subsequently regrouped into two units (1–3 and 4–7), and results of the analysis revealed a high number (101 of 149) of correctly classified observations (PCC = 67.79; c value $< .001$, 1000 trials). The multigram in Figure 9.5 shows the correctly and incorrectly classified observations and again reveals much room for improvement if a causal explanation is to be advanced. Nearly half of the persons (29 of 59) with low levels of fatty acids in their blood nonetheless consumed grilled meat 4–7 days per week. Moreover, 19 persons with high levels of fatty acids in their blood consumed grilled meat only 1–3 days per week. These more equivalent frequencies of units may explain the lower CSI values (ranging from .87 to .91) for this analysis. In any event, the incorrectly classified observations must eventually be accounted for in future work, and of course this study must be exactly replicated to gain initial evidence for the generality of its conclusions.

Mediation models can therefore be analyzed with observation oriented modeling, but it requires a deeper commitment to thinking about how the causal forces operate through the ordering of the observations. Unlike variable-based modeling, in which scale differences are usually removed through standardization, observation oriented modeling requires an explanation of differing numbers of units in the ordered observations, which again attests to the importance of developing an integrated model. The analysis also works from the notion of effects conforming to causes and therefore schematically works from right to left in an ordered chain of

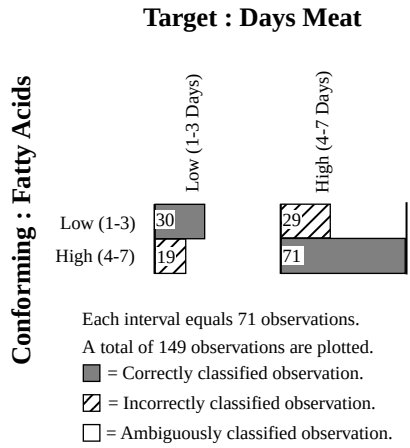

Figure 9.5 Multigram conforming fatty acids (two units) to consumed grilled meat (two units)

observations. Only observations that are classified correctly are carried from one step in the analysis to the next as the researcher works his or her way back through the complete chain of ordered observations. This rule may seem restrictive, particularly given the generally noisy nature of psychological data. It is easy to imagine a great deal of attrition as one works back through the causal chain, but this is simply the price that must be paid when realizing that causes are found only at the level of the observations (usually persons) in the model, not at the level of the aggregate. It is senseless in the previous example to speak of the mean level of grilled meat consumption causing the mean blood level of fatty acids; rather, it is only valid to describe how the meat consumed by an individual initiates particular chemical reactions that result in an increase in his or her fatty acids. These fatty acids are then implicated in the cancer that may be found in any given person's pancreas.

PERSON-CENTERED MEDIATION ANALYSIS

The fact that causes operate only at the level of the persons (or observations) in a study has prompted a number of methodologists to devise different techniques for addressing mediation models. Collins, Graham, and Flaherty (1998) departed from the traditional variable-based approach to mediation analysis by first making it clear that a mediation model represents a chain of events connected as efficient causes: "According to our definition, in its

strictest form a mediated process is analogous to a line of dominos. Knocking over the first domino starts a sequence where the rest of the dominos are knocked over one after another" (p. 297). They go on to state very clearly that these causes operate at the level of the persons (or observations) in the study:

> *Another implication is that the mediated process is, in its entirety, an intra-individual phenomenon. In other words, it is the individual who undergoes the process of the independent variable leading to the mediator, and the mediator then leading to an outcome.*
>
> ***(p. 297)***

Consistent with this argument, Collins and colleagues note that a necessary condition for demonstrating mediation is the longitudinal assessment of persons at a minimum of three different time points. With generic variable-based mediation modeling, it is considered permissible to assess variables cross-sectionally and analyze the aggregated data following scripted steps that ostensibly demonstrate what are known as partial or full mediation effects. To demonstrate the differences of modeling mediation at an individual and aggregate level, Collins and colleagues present a data set that demonstrates full mediation at the level of the aggregate but not at the level of the individuals. The mediation model in Figure 9.6 shows two efficient causes linked in time. The first is a drug abuse prevention program, and in the language of observation oriented modeling it is composed of two units: treatment or control. The second efficient cause, which is initiated by the first efficient cause, is observed in two units as well: trained to resistance of drug offers or not trained to resistance. It is interesting that Collins and colleagues discuss the notion of a "threshold," much like the pancreatic cancer example, when justifying the dichotomous nature of these ordered observations. Although they certainly did not adopt the perspective of observation oriented modeling, this perhaps represents a tacit acknowledgment of the necessity of similarly structured observations when modeling efficient causes. The final effect in the model is also ordered in two units as resisting an actual drug offer or failing to resist a drug offer.

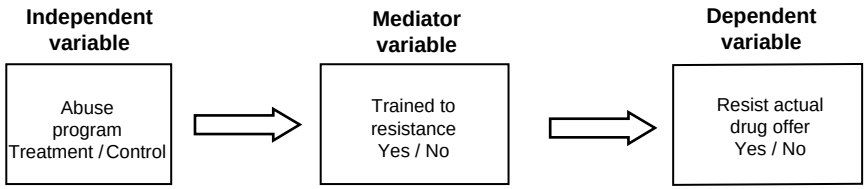
Independent variable
Abuse program Treatment / Control
Mediator variable
Trained to resistance Yes / No
Dependent variable
Resist actual drug offer Yes / No

Figure 9.6 Person-centered mediation model for drug abuse program

Analyzing the observations for the model shown in Figure 9.6 via the traditional variable-based approach yields results consistent with a full mediation effect, which is interpreted to mean that (1) the independent variable is correlated with the mediator and (2) when analyzed together in multiple regression, the mediator accounts for a statistically significant portion of variance in the dependent variable, whereas the independent variable is not significantly related to the dependent variable. In a sense, at the level of aggregated observations (means and covariances), the independent variable is superfluous in the context of the mediator, and this is an ideal outcome for the psychologist positing a mediation model. Consistent with the techniques employed by Collins and colleagues, however, observation oriented modeling reveals a different interpretation of the data because it is not based on aggregates.

As was previously shown, the chain of observations is analyzed sequentially working backwards from the last effect in the sequence. The resistance observations (resist drug offer; fail to resist) are therefore first rotated to conformity with the training observations (trained to resistance; not trained), both of which are ordered as two units. The results of the analysis are fairly impressive, as shown in the multigram in Figure 9.7. A large number of the persons ($n = 147$) who were not trained to the threshold level of resistance later accepted an actual drug offer, whereas 18 persons trained insufficiently successfully refused a drug offer. With regard to those

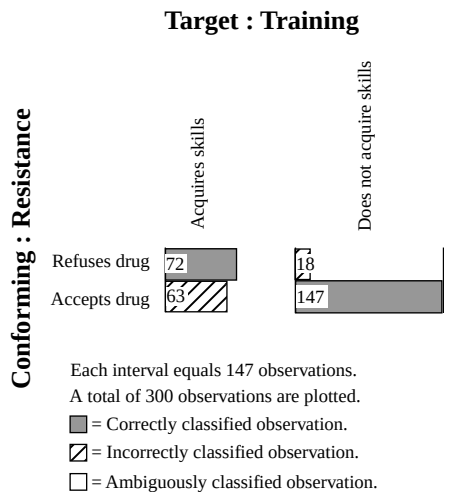

Figure 9.7 Multigram conforming responses to a drug offer to training of resistance to such offers

individuals who were trained to and beyond the threshold, the resistance level was more evenly split, with 72 persons refusing a drug offer and 63 accepting. Overall, the pattern was nonetheless consistent with expectations regarding the efficacy of training: PCC = 73.00% and c value < .001 (1000 randomized trials).

For the second step in the analysis, the 219 correctly classified persons of the original 300 (73%) were identified. The second efficient cause, training to resistance, which was also seen as the effect of the original cause, was then rotated to conformity with the treatment/control ordered observations. The PCC value (64.38%, 141/219) exceeded 50%, and the c value was low (c = .001, 1000 trials). Examination of the multigram in Figure 9.8, however, reveals mixed results regarding the expected causal pattern. It is important to keep in mind that the observations shown in the multigram are for those persons who were classified correctly in the first step of the analysis. This means that the patterns of observations for these persons were consistent with a causal understanding of the training sessions as they impacted the acceptance or refusal of an actual drug offer. They either trained to the threshold and resisted the drug offer or they did not train to the threshold and accepted the drug offer. Figure 9.8 shows that the majority of persons in the control group (105 of 141, 74%) did not acquire the necessary skills and then later accepted a drug offer. This result is consistent with the causal understanding of the observations. Unfortunately, a slight majority of the persons in the drug treatment condition (42 of 78, 54%) also failed to acquire the resistance skills and later accepted a drug offer,

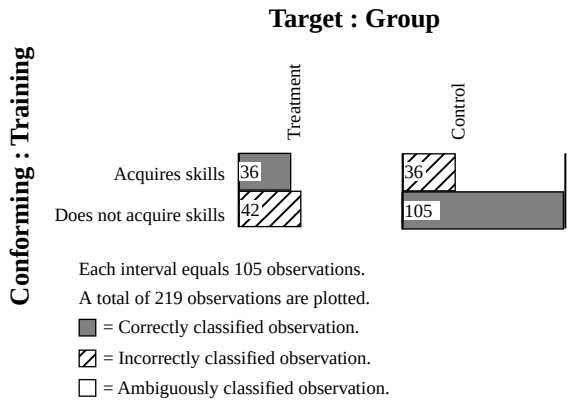

Figure 9.8 Multigram conforming training of resistance to drug offer to group membership. Only persons successfully classified in Figure 9.7 are included.

contradicting the causal expectations of the observations. The overall strength of classifications for the analysis was also relatively low, with CSI values ranging from .78 to .90.

The observation oriented analysis therefore shows that when causes are properly understood to operate at the level of the observations—in this example, at the level of the persons—results quite different from variable-based modeling can be found. "Full" and "partial" mediation in the variable-based approach only have meaning when understood in terms of explained variances and covariances (aggregate statistics) between the three variables in the mediation model. The terms do not indicate the extent to which a cause-and-effect sequence can be completely or partially traced through a series of observations. Full mediation for the drug treatment data was observed for 141 of 300 persons. A total of 105 persons in the control group did not acquire the necessary skills to resist a drug offer and in fact later accepted such an offer, and 36 persons in the treatment group acquired the skills to resist a drug offer and later successfully refused an offer. Unfortunately, observations for 159 persons (see Figures 9.7 and 9.8) were inconsistent with expectations based on an understanding of the treatment and training causal influences on resisting offered drugs. For instance, 63 persons were trained to criterion but still accepted a drug offer (see Figure 9.7), and 42 persons who accepted a drug offer and had not acquired the necessary skills for resisting such an offer had nonetheless been assigned to the treatment group (see Figure 9.8).

Alexander von Eye (2002) similarly developed a person-centered approach toward mediation referred to as configural frequency analysis (CFA). Like the approach of Collins and colleagues (1998) discussed previously, as shown in the observation oriented analyses, von Eye's method depends on categorizing observations into dichotomous groups. The different configurations, or patterns of observations, are then examined. For instance, the three sets of ordered observations for the drug intervention data presented previously are first crossed, creating nine groups (e.g., treatment/training to threshold/refuse drug offer and control/training below threshold/accept drug offer). The frequencies for the different groups are then examined and interpreted in light of the mediation model. Unlike the approach of Collins and colleagues, von Eye's method does not require longitudinal data and may thus distort the efficient cause nature of mediation models. By focusing on patterns of observations that are not necessarily ordered in time, CFA borders on addressing formal causes rather than efficient causes. Regardless, one way

to adapt observation oriented modeling to von Eye's method is to use logical hypothesis testing or to cross ordered observations in the modeling process. The following expression, for instance, can be tested for the drug treatment observations in the Observation Oriented Modeling software:

$$\text{Group} \rightarrow \text{training} \wedge \text{resistance},$$

where "^" indicates that the training and resistance ordered observations have been fully crossed prior to being rotated to conformity with the group ordered observations. Results for this expression were impressive in terms of the PCC index (71.00%, 213/300) and c value (<.001, 1000 trials), but the multigram in Figure 9.9 shows the same mixed findings described previously. On the one hand, observations for 105 persons were correctly classified and fit the expected casual sequence: No treatment (control) → does not acquire skills → accepts a drug offer. An additional 36 persons were correctly classified and fit the following sequence: Treatment → acquires skills → refuses drug offer. Unfortunately, an equal number of persons ($n = 36$) from the control group fit a similar pattern: No treatment → acquires skills → refuses drug offer. The observations for 123 other persons were also inconsistent in different ways with the expected sequence of observations based on an efficient cause understanding of their relationships. von Eye's approach therefore

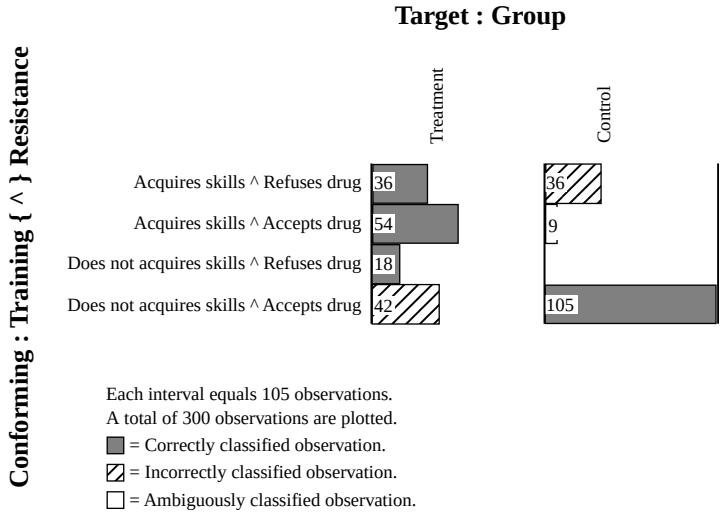

Figure 9.9 Multigram conforming training and resistance to group membership

leads to the same conclusions as those presented previously in which the observations yield mixed results. At the level of the persons in the study, some evidence is found for the efficient cause model, but most of the observations violate the expected patterns. This interpretation stands in stark contrast to the deceptive conclusion of "full mediation" from the traditional variable-based analysis.

LOGICAL HYPOTHESIS TESTING

The person-centered methods of mediation analysis suggest that causality may sometimes be approached from the perspective of logic and set theory. Indeed, in the most generic sense, "a cause is that which, being present, the effect is there, and being removed, the effect is taken away" (Wallace, 1996, p. 269). Although this definition stresses a logical understanding of causality, it in no way excludes Aristotle's four species of cause nor Mill's methods. Considering the cause as logically necessary for the effect is particularly propitious for detective-like thinking that works from the effect back to the cause. Consider a contrived example in which a restaurant owner begins with a clear effect, namely 10 of his 38 customers from the previous evening suffered from food poisoning. What is the cause of the food poisoning? Based on his knowledge of modern science, the owner is confident a microbe (formal, material causes) with capabilities (formal, efficient causes) of producing illness in humans (final cause) must be involved, but what is the carrier of the microbe? The effect, food poisoning, is dependent not just on the microbe but also on the food that carried the microbe to the digestive tracts of the people who became ill. The effect is also dependent on other simultaneous causes or conditions, such as the persons' immune systems being unable to combat the microbe. Ignoring these other conditions for the current discussion, however, and following Thomas Aquinas and the Schoolmen, the microbe and food can be differentiated with respect to their understood causal implication; namely, the presence of both the food and the microbe produced the effect, but the microbe is clearly the principal cause, whereas the food is an instrumental cause.[10] The owner's search is thus for

[10] St. Thomas Aquinas gives an example of an instrumental cause when describing a minister (the instrumental cause) delivering the Sacraments in which God is the principal cause of the change brought about (*Summa Theologica*, 3a, 64, 1–2). For a more general discussion of causality as understood by Aquinas and the Scholastics, see Harper (1940).

the instrumental cause through which the principal cause achieved its effect. The foods in the restaurant are normally free of harmful microbes and therefore do not normally cause illness. Some foods (formal, material causes), however, are capable (formal, final causes) of carrying (efficient cause) particular pathogenic microbes and were consumed by people in the restaurant. If these foods were not capable of carrying and sustaining the lives of the microbes, the effect would not have been observed. Which of the foods (chicken sandwich, hamburger, salad, French fries, or shake) was contaminated and consumed by those people who became ill? The restaurateur decides to examine all the receipts of his customers and search for the food that was consumed by each person who became ill and was not consumed by the persons who did not become ill.

Observation oriented modeling provides an efficient means for accomplishing this goal. All of the observations are composed of two units (Figure 9.10). Each customer either consumed or did not consume each of the foods and then either did or did not become ill. The effect (food poisoning) is simply rotated to conformity with each of the five potential instrumental causes (chicken sandwich, hamburger, salad, French fries, or shake) in the Observation Oriented Modeling software. The results show that two foods are implicated in the food poisoning: the hamburger (PCC = 100, 38/38; c value $< .001$, 1000 trials) and the French fries (PCC = 73.68, c-value = .001). The PCC values for the other foods were all below 61%, with c values greater than .49. Given that all of the persons who ate hamburgers grew ill and those who did not eat hamburgers remained healthy, the smoking gun seems to be the hamburger. But what of the French fries? Knowing that hamburgers are usually served with French fries, the persons who ate French fries with or without a hamburger become particularly important. Those persons can simply be selected and the incidence of food poisoning noted. In this example, none of the seven people who ate French fries without a hamburger contracted

ID	Sandwich	Burger	Salad	Fries	Shake	Poisoned
1	0	0	0	1	0	0
2	0	1	1	1	1	1
3	0	1	0	1	0	1
4	1	0	1	1	0	0
5	0	1	1	1	0	1
6	0	1	0	0	0	1
7	1	0	1	1	0	0

Figure 9.10 Example restaurant observations for seven individuals

the illness, whereas all eight of the people who ate French fries with a hamburger became sick. This pattern of results clearly leaves the hamburger as the lone instrumental cause.

The analysis could also have been addressed using compound logical statements. Such statements abound in every person's day-to-day thinking, and they are quite common in the natural sciences. Medical diagnosticians, for instance, utilize these sorts of statements as a matter of necessary habit. A sore throat without fever would indicate a common cold, whereas a sore throat with fever would suggest influenza. Bright red blood in the stool may indicate irritated hemorrhoids, whereas blackish blood and abdominal pain may indicate a more serious condition such as colon cancer. A yellowish skin tone, yellowing of the eyes, and pain in the middle portion of the back but not in the spine may indicate pancreatic cancer. These examples represent different compound logical statements regarding various effects that may indicate a single principal cause:

$$p_{\text{sore throat}} \wedge \sim q_{\text{fever}} \equiv r_{\text{common cold}}$$

$$p_{\text{bright red blood}} \wedge \sim q_{\text{black blood}} \wedge \sim r_{\text{abdominal pain}} \equiv s_{\text{hemorrhoids}}$$

$$p_{\text{yellowish skin tone}} \wedge \ q_{\text{yellow eyes}} \wedge \ r_{\text{middle back pain}} \wedge \sim s_{\text{pain in spine}}$$

$$\equiv t_{\text{pancreatic cancer}}.$$

Clinical psychologists who adopt the medical model for diagnosis regularly rely on these types of statements as well, but they have no flexible and convenient way to evaluate or test them within the Pearsonian–Fisherian framework. Each observation represents the noted presence or absence of some attribute, and this complete lack of continuous quantities of course rules out parametric techniques. Nonparametric statistics based on randomization tests (to avoid testing abstract population parameters) could possibly be used, but in variable-based modeling the observations are not easily combined to form logical compound statements.

In observation oriented modeling these limitations are overcome because the assumption of continuous quantity is not made, and as shown in Chapter 2, observations converted to their deep structure can easily be combined to match logical statements. Consider the example observations shown in Figure 9.11. Here, the effect is even more serious than that faced by the restaurant owner: 11 people have committed suicide. Based on reports from friends and family members, the psychological researcher determined the presence or absence of different events in the 11 persons'

ID	Physical Abuse	Emotional Abuse	Recent Loss	Spiritually Active	Clinical Depression	Suicide
1	0	0	0	0	1	0
2	0	0	1	0	0	0
3	0	1	0	0	1	1
4	0	0	0	0	1	0
5	0	1	0	1	1	0
6	0	1	0	1	0	0
7	0	1	0	0	1	1

Figure 9.11 Example suicide observations for seven individuals

lives as well as the lives of 31 comparison persons; namely, childhood physical abuse, childhood emotional abuse, a recent loss (e.g., death of a loved one or loss of a romantic relationship), clinical levels of depression, and activity in a spiritual organization such as a church, synagogue, or prayer group. Ideally, an integrated model would drive or organize such observations, but beginning generically, the goal may be to establish a formal cause model of suicide composed of a logical profile. For instance, perhaps the profile that discriminates between those who committed suicide and those who did not is of the form,

$$p_{\text{physical abuse}} \vee q_{\text{emotional abuse}} \vee r_{\text{depression}} \wedge s_{\text{loss}} \wedge \sim t_{\text{spiritual support}} \equiv u_{\text{suicide}},$$

or equivalently,

$$\sim p_{\text{physical abuse}} \wedge \sim q_{\text{emotional abuse}} \wedge \sim r_{\text{depression}} \wedge \sim s_{\text{loss}} \vee t_{\text{spiritual support}} \equiv \sim u_{\text{suicide}}.$$

In other words, the persons who committed suicide suffered from physical abuse or emotional abuse or clinical depression, and they suffered a loss and did not have spiritual support. The logical antithesis is that the persons who did not commit suicide did not suffer from physical abuse, emotional abuse, and depression, or they did not suffer a loss or they had spiritual support. Using the deep structure of the observations, such logical statements can be created as novel sets of observations ordered into two units as "true" or "false." These observations can then be rotated to conformity with the suicide observations ordered as "yes" or "no." Generating observations for the first statement in the Observation Oriented Modeling software and performing the rotation yielded unimpressive results (PCC = 40.48, 17/42; c value = .65, 1000 trials). If each set of prior history observations is analyzed separately and conformed to the suicide observations, only physical abuse is causally implicated in the

suicides (PCC = 78.57, 33/42; *c* value = .01, 1000 trials). The results for physical abuse further show that 6 of the 11 people who committed suicide were correctly classified, whereas 27 of the 31 people who did not commit suicide were correctly classified. Particularly given that only 55% of the persons who committed suicide were correctly classified, the model needs improvement.

Can any of the other observations regarding the persons' histories be used to explain suicide? If answering this question is regarded as a search for causes, then an integrated model must be created to explain the connections between the ordered observations. It may be the case, for instance, that spiritual activity operates as an accidental cause in conjunction with one or more of the other sets of ordered observations. Aquinas (2008) describes an accidental cause as "either one that removes an impediment, or the removal itself of an impediment" (*Summa Theologica*, 1a 2æ, 76, 1; see also, 1a 2æ, 75, 1). Applying this principle to the current observations, spiritual activity may act as a sort of causal "buffer" against suicide in the context of other causes. It may further be argued that emotional abuse often accompanies physical abuse, and that emotional abuse may simply be more common, even with the persons who did not commit suicide. Consequently, the emotional abuse observations were not by themselves conformable to the suicide observations, but perhaps in logical conjunction with the physical abuse observations, their causal impact might be revealed. Putting these ideas together yields the following logical compound statement:

$$p_{\text{physical abuse}} \vee q_{\text{emotional abuse}} \wedge \sim t_{\text{spiritual support}} \equiv u_{\text{suicide}}.$$

Conforming the ordered observations generated from this statement to the suicide ordered observations yielded remarkable results: PCC = 97.62,

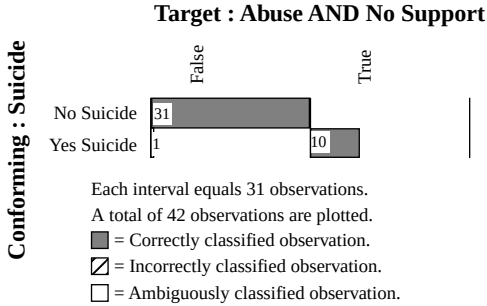

Figure 9.12 Multigram conforming logical profile to suicide observations

41/42; c value $< .001$, 1000 trials. The multigram in Figure 9.12 shows that 10 of 11 persons who committed suicide fit the profile entailed in the logical statement. The one person who committed suicide and was misclassified did not suffer from reported abuse and was not spiritually active.[11]

FINAL, FORMAL, AND EFFICIENT CAUSES

The previous examples of logical modeling demonstrate reasoning from the effect, suicide, to the cause or causes, such as childhood physical abuse. From Aristotle's perspective, this is often the first step taken en route to *epistēmē* because it is normally the case that effects are initially known better than their causes. Through careful observation, experimentation, and dialectical reasoning, the causes can be made well-known, and then reasoning can proceed from causes to effects and presented in the form of a demonstrative syllogism such as that shown previously. At that point, scientific knowledge of nature (*scientia naturalis*) is said to be possessed. William Wallace (1996) refers to this progression as the "demonstrative regress" and evidences its application in numerous

[11] It might further be asked if any logical combination of only two sets of observations performs as well as all three ordered observations. In observation oriented modeling, this type of question is answered by dropping one set of observations from the logical statement. For instance, the spiritual activity observations can be dropped:

$$p_{\text{physical abuse}} \vee q_{\text{emotional abuse}} \equiv u_{\text{suicide}}.$$

The suicide ordered observations can then be rotated to conformity with the observations from this simpler compound statement and the PCC value (69.05%) noted. A c value can be computed by comparing the randomization results from the models for the original compound logic statement and the simpler compound statement. Without spiritual activity in the compound statement, the PCC value for the original model exceeded the PCC value for this model for 732 of the 1000 randomized trials, $c = .268$. The increase in the PCC value from the simpler logical compound statement (69.05%) to the original statement (97.62%) was only somewhat unique. Removing either of the abuse observations from the original logical statement resulted in less substantial declines in the PCC values (88.10 and 83.33%) and c values of approximately .50. Using either physical abuse or emotional abuse observations in combination with the spiritual activity observations therefore yields impressive results that are not remarkably different (as indicated by the PCC and c values) from the original compound statement. Nonetheless, the original logical statement produced the highest PCC value, and exact replication must be the final arbiter of its accuracy compared to simpler statements.

historical and contemporary examples in the biological and physical sciences.[12] The outcome of the demonstrative regress is an integrated model that sufficiently explains some aspect of nature. Although no models were presented for the previous example, they would be necessary for scientific knowledge of suicide. In other words, accuracy in classifying the observations is not in itself sufficient; a model that explicates causes and effects must also be developed. It is through such models that the fruits of the scientist's inductive and deductive work, as well as his or her illative sense, come together to explain the largest and smallest things of the cosmos and that enigmatic microcosm, the human person.[13]

Two particular features appear to separate humans from all other creatures in the universe: the capability of extensive self-reflection and the capability to construct complex mental images of future events and then behave for the sake of such visions. Goal-directed behavior falls under the philosophical consideration of teleology and is essentially what Aristotle had in mind when discussing final cause regarding human action. For Aquinas, final cause was in fact the most important cause, both theologically and psychologically:

> *I answer that, every agent, of necessity, acts for an end. For if, in a number of causes ordained to one another, the first be removed, the others must, of necessity, be removed also. Now the first of all causes is the final cause. The reason of which is that matter does not receive form, save insofar as it is moved by an agent; for nothing reduces itself from potentiality to act. But an agent does not move except out of intention for an end. For if the agent were not determinate to some particular effect, it would not do one thing rather than another:*

12 William Wallace describes this demonstrative regress in *The Modeling of Nature* (1996, pp. 300–308). William Harvey, who is best known for his explanation of how blood circulates through the human body, described this demonstrative regress in the introduction of his work *Anatomical Exercises on the Generation of Animals* (1651/1952): "Since in animal generation (and, indeed, in all other subjects upon which information is desired) inquiry must be begun from the causes, especially the material and efficient ones, it appears advisable to me to look back from the perfect animal, and to inquire by what process it has arisen and grown to maturity, to retrace our steps, as it were, from the goal to the starting place; so that when at last we can retreat no farther, we shall feel assured that we have attained to the principles" (p. 335).

13 "Now, the object of an unprejudiced psychology is the complex being whom we call *Man*. Medieval philosophers said he was a *microcosm*, to indicate how far his wealth of activities is a kind of synthesis of all cosmic energies" (Mercier, 1918, p. 218).

> *consequently in order that it produce a determinate effect, it must, of necessity, be determined to some certain one, which has the nature of an end.*
>
> **(Summa Theologica, 1a, 2æ, 1, 2)**[14]

In modern psychology, teleology is not taken seriously, at least not to the point of being considered as central to our understanding of human action. It is instead treated as a topic of special consideration under goal research, and even these modest considerations represent an improvement over previous years. As adroitly researched and explained by Joseph Rychlak, both formal and final causes have been implicit in various psychological theories throughout the years, but these causes have not been made explicit nor played a conscious role in theory development. Consistent with an Aristotelian viewpoint, Rychlak argues that scientific knowledge of persons will be limited until psychologists incorporate these forgotten causes into their theorizing.[15] Following his own advice, Rychlak (1988) incorporated formal and final causes into his logical learning theory, which he constructed and tested over the course of 20 years. Teleology plays a leading role in this theory, which was created as an alternative to behavioral learning theories, as can clearly be seen in the unique notion of a *telosponse* as

> *the person's taking on (premising) of a meaningful item (image, word, judgmental comparison, etc.) relating to a referent acting as a purpose for the sake of which behavior is then intended… When the individual behaves "for the sake of" this purpose he is telosponding or acting intentionally.*
>
> **(p. 283)**

Other psychologists have similarly incorporated Aristotelian notions of cause in their theorizing, although not always acknowledging the four causes. Walter Mischel (2004) wrote of a science of personality based on his Cognitive–Affective Processing System that seeks to integrate the cognitive, affective, and motivational forces at work within the individual. Mischel wrote of exploring the patterns (formal causes) within a person's cognitive schema, understanding how a person pursues goals (final causes),

14 Aquinas also makes this point in Book III of his *Summa Contra Gentiles* (1975): "For every being is ordered to its end through its action" (p. 70), and "Moreover, the end holds first place over other types of cause, and to it all other causes owe the fact that they are causes in act: for the agent acts only for the sake of the end, as was pointed out" (p. 73).

15 This is not to suggest that Rychlak subscribes to philosophical realism in the tradition of Aristotle and Aquinas. He famously frames the battle lines between Lockean and Kantian philosophy and sides with the latter. His extraordinary personality textbook is organized around this dialectic (Rychlak, 1981).

and mapping the different "if … then" cognitive relationships (efficient causes) that explain why people behave consistently within similar situations. Similarly, Dan Cervone (2004) introduced his Knowledge-and-Appraisal Personality Architecture model to explain the cross-situational coherence in patterns (formal causes) of responses to different stimuli, usually judgments of self and others. Cervone particularly draws upon the realist philosophy of Rom Harré, emphasizing the importance of understanding the structures (formal causes) and processes (efficient causes) that constitute personality.

A common ancestor to Rychlak, Mischel, and Cervone is George Kelly, who included final cause in both the central philosophical premise and the fundamental postulate of his personal construct theory. With regard to philosophy, Kelly (1955) described constructive alternativism as a person's capacity to create different constructions of the world: "We assume that all of our present interpretations of the universe are subject to revision or replacement" (p. 15). His fundamental postulate states, "A person's processes are psychologically channelized by the ways in which he anticipates events" (p. 46). Understanding how a person goes about anticipating the future is thus of utmost importance, a view that fits well with Aquinas' placing final cause as first among the causes.[16] Formal cause plays a prominent role in personal construct theory as well with regard to how people construe the world. As described in Chapter 8, people are held to construe the world through bipolar personal constructs that are arranged in hierarchies. Kelly is here relying on formal cause to explain how the inner conceptual realm of the psyche is structured and ordered. He consequently viewed psychotherapy as a process of helping the client to reorganize his or her personal construct system (formal cause) in order to anticipate the future (final cause) in alternative and more productive ways. The dilemma faced by Kelly, his descendants, and others who wish to develop such models, however, is that the Pearsonian–Fisherian approach toward study design and data analysis is fundamentally inconsistent with a person-centered, realist view of causality. Examples in this chapter and previous chapters show that observation oriented modeling provides a framework and set of tools that is entirely consistent with these endeavors.

[16] This may be the extent of any overlap between Thomas Aquinas and George Kelly. The latter is considered one of the fathers of modern constructivism, and he did not use the thought of Aristotle or Aquinas at all in the development of his personal construct theory.

	Self	Ideal Self	William	Dad	Mom	Patrick	Julie	
Masculine	0	1	-1	1	-1	1	-1	Feminine
Educated	1	1	-1	1	-1	1	1	Uneducated
Uncomfortable with self	-1	-1	-1	-1	-1	-1	-1	Comfortable w/self
Honest	1	1	1	-1	1	1	-1	Dishonest
Conventional	1	-1	-1	1	1	0	1	Creative
Scholarly	-1	1	1	-1	1	1	1	Unscholarly
Intelligent	1	1	1	-1	1	1	-1	Unintelligent
Scientific	0	1	0	-1	1	0	0	Unscientific
Unidimensional	1	-1	-1	1	-1	-1	-1	Multidimensional
Religious	-1	1	-1	1	1	1	1	Anti-religious

Note. 1 = Construct Pole on left applies; -1 = Construct Pole on right applies; 0 = Neither Pole applies to the person

Figure 9.13 Repertory grid ratings for a hypothetical therapy client

One method that might prove useful in addressing questions of final and formal cause involves the testing of two- and three-dimensional patterns. For example, consider the repertory grid shown in Figure 9.13, in which a client judged himself, his ideal self, and five other people on 10 of his own personal constructs (i.e., descriptive terms that he provided during a structured interview). From Kelly's (1955) perspective, these personal constructs represent a sample of the bipolar dimensions through which this particular person makes sense of himself and the world. Repertory grids are often analyzed in ways that reveal the overall pattern or patterns in the ratings. For instance, the singular-value decomposition plot in Figure 9.14 shows that the client construes his ideal self and the other five people as generally unlike himself. In other words, the pattern represents what clinical psychologists might refer to as "self isolation" (Norris & Makhlouf-Norris, 1976).

Using the tools of observation oriented modeling, at least two analyses that coincide with Figure 9.14 can be conducted. First, the pattern of judgments for the client's self can be compared to the pattern for his ideal self. Second, the pattern of judgments for the client's self can be compared to the pattern of judgments for the other five people in the repertory grid. Figure 9.15 shows that the pattern of ratings for the client's self across the 10 constructs does not generally match the patterns of ratings for his ideal self and the five other people in the repertory grid. In fact, he construed himself as similar to his ideal self on only 2 of 10 constructs (20%, c value = .89, 1000 trials), and he infrequently applied the same poles of the

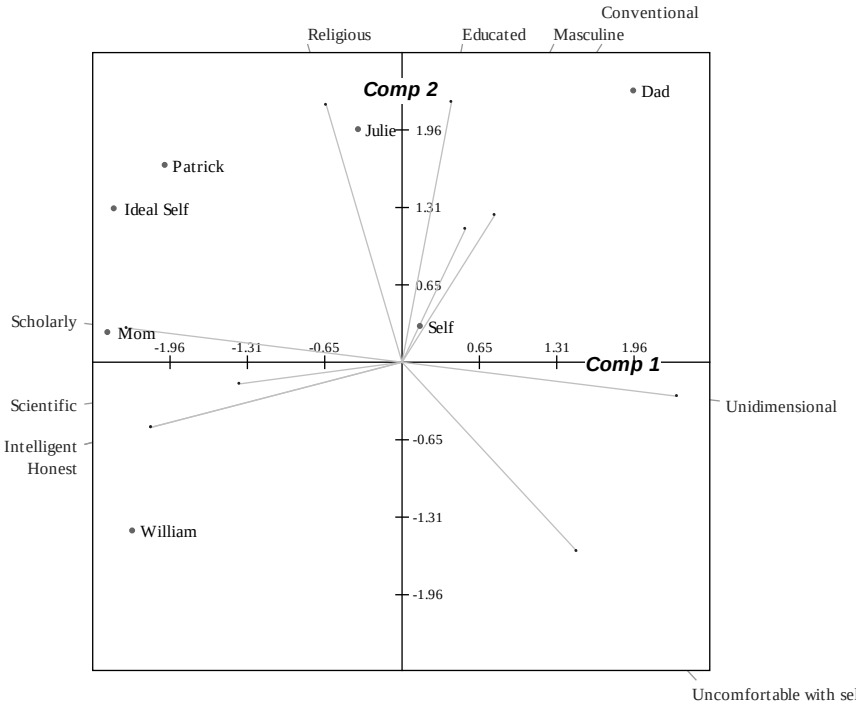

Figure 9.14 Singular-value decomposition plot for repertory grid ratings

10 constructs to both himself and the other people (34%, *c* value = .51, 1000 trials).[17] With such patterns in the application of the client's constructs, the goal of therapy would be to help him reconsider his current constructions or perhaps to develop new constructs that are less isolating. In this client's case, it was difficult for him to change how he applied most of his current constructs to himself and the other people. Only a few old constructs were thus retained, and an effort was made to discover or create different

[17] The analysis comparing the self ratings to the other ratings was conducted in the Observation Oriented Modeling software by first transforming the ratings to their deep structure. The deep structure of the self ratings was then compared to the deep structures of the ideal self and other ratings, and the number of exact matches was tallied. The *c* value was computed by randomizing the deep structures of the ideal self and self ratings and tallying the matches and comparing the results to the original number of matches for the actual observations. In this way, the Observation Oriented Modeling software can be used to examine patterns of observations with regard to how well they match an expected pattern. The *c* value again indicates how unusual the match is between the expected and observed observations in their deep structure forms.

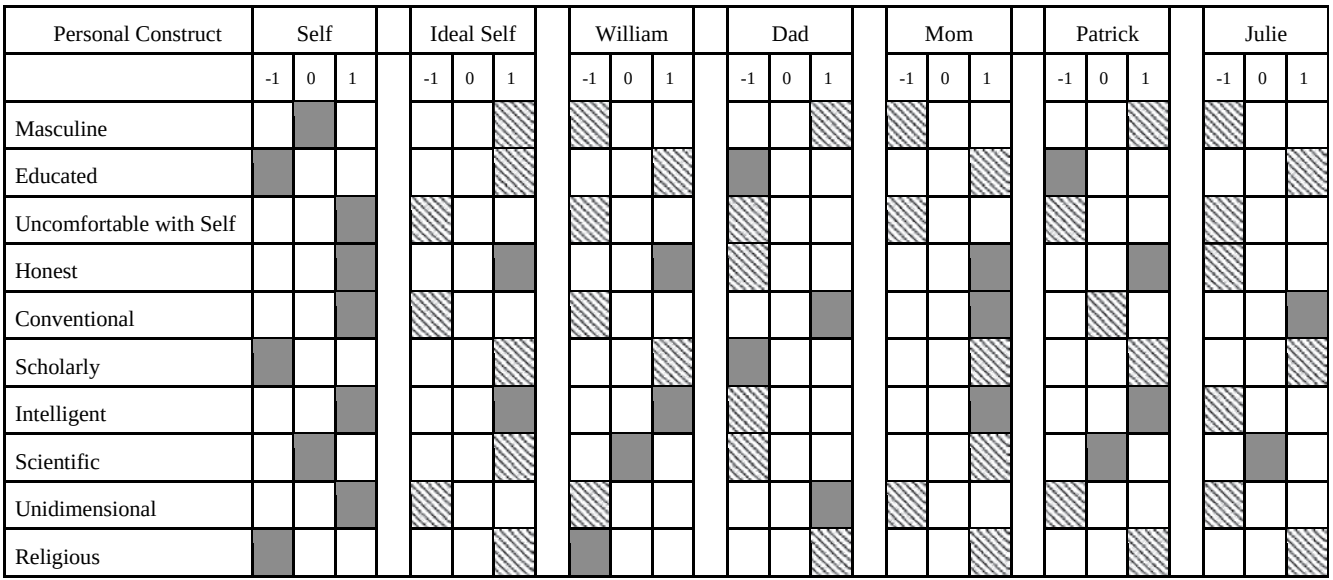

Figure 9.15 Repertory grid ratings presented as patterns in which gray boxes indicate matches with the self and diagonal boxes indicate mismatches

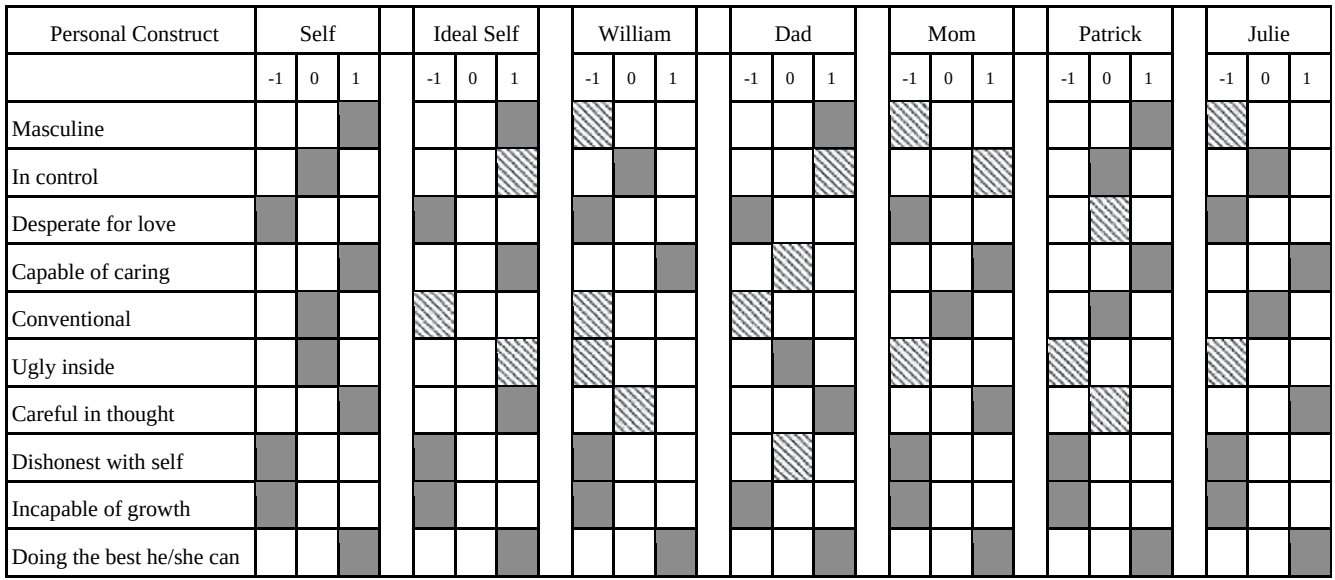

Figure 9.16 Second repertory grid ratings presented as patterns in which client shows greater similarity between self and others

constructs to help the client creatively construe the world in novel ways. Figure 9.16 shows the pattern of responses from a second grid obtained near the end of the client's involvement in therapy. The new patterns show that the client construes himself as similar to his ideal self on 7 of 10 constructs (70%, c value = .018, 1000 trials), and he applies the same poles of the constructs to both himself and the other people with greater frequency (68%, c value < .001, 1000 trials). The differences in the proportions of matches between the pre- and post-therapy repertory grids were also highly unusual compared to randomized versions of the ratings (c values < .02). This example shows one way in which final and formal causes can be utilized and evaluated. The goal of bringing the self and ideal self into agreement on a set of constructs is the final cause, and the pattern of construing revealed in the repertory grid ratings is the formal cause.

TRANSITION TO THE FOUR CAUSES

Cognitive behavioral therapists and other psychologists who do not normally rely on Kelly's personal construct theory often employ methods similar to those discussed previously and can hence relate well to the pattern matching analyses.[18] Indeed, the transition to observation oriented modeling for most psychologists may begin simply with a re-examination of existing models, even if the models are represented purely in variable form. For example, in Chapter 8, the following regression model expressing relationship commitment as an additive function of relationship satisfaction, investment, and attractive alternatives was shown:

$$\text{Commitment} = a + b_1\,\text{satisfaction} + b_2\,\text{investment} - b_3\,\text{alternatives} + e.$$

In variable-based language, commitment is the dependent, or outcome, variable, whereas satisfaction, investment, and alternatives are the

[18] An example in which clinical psychologists investigated the structural aspects of clients' constructions can be found in Dozois et al. (2009). Note, however, that the researchers "measured" structural changes using a summary statistic based on numerous ratings that reflected their overall discrepancy or similarity. They did not assess patterns, per se, of ratings as is necessary when discussing changes in structure. It is also noteworthy that the researchers claimed to "test the cognitive mechanisms underlying the efficacy of CT [cognitive therapy]" (p. 1080), but they presented no integrated model to expound exactly how such causes operate at the level of the individuals in their study. The claim for establishing cause was backed by the examination of average differences between groups in a randomized controlled experiment.

independent, or predictor, variables (Rusbult, 1980). The theory behind the regression model justifies the predictors as ingredients that go into the creation of a committed relationship. Although this reasoning may suggest a formal cause model, it is not described as such and is instead largely construed pragmatically as a means for predicting the longevity of romantic relationships. Reasoning from Aristotle's four causes, however, the model can be construed differently. A person committed to his or her relationship necessarily imagines a future with the other person. This plan represents a clear intention, or purpose, and can be considered as a final cause. Example questions used by Caryl Rusbult and colleagues to assess commitment clearly demonstrate a future orientation:

> For how much longer do you want your relationship to last?
>
> Do you ever have fantasies about what life might be like if you weren't dating your partner?

In contrast, questions pertaining to relationship satisfaction do not reveal future intentions but instead invoke reflections on the current status of the relationship:

> All things considered, to what degree do you feel satisfied with your relationship?
>
> How much do you love your partner?

Without an integrated model from which to work, it is difficult to judge whether such statements should reflect formal causes (representing meaningful acts of self-construal) or perhaps effects of relationship involvement. The questions meant to assess investments and alternatives reveal both future intentions and contemporaneous self-predications:

> Have you invested things in your relationship that you would in some sense lose if the relationship were to end?
>
> How much have you got invested in your relationship?
>
> If you weren't dating your partner, would you do okay—would you find another appealing person to date?
>
> How do your alternatives compare to your relationship with your partner?

From the perspective of observation oriented modeling, each of these statements must be classified as a cause of some type or as an effect in an integrated model. The procedures described in Chapter 2 can then be used to rotate the effects to maximal agreement with their causes. The central issue to be confronted by the relationship researcher is thus causality rather than pragmatic prediction using a variable-based model. It is interesting to note as well that with observation oriented modeling, careful attention is

paid to the particular questions being asked and the self-reflections they are presumed to invoke in the respondents. In light of modern psychometric theory, each item is considered as an error-laden indicator of a continuously structured quality, or latent trait. Reflecting this assumption, people are asked to respond to each of the previous questions using an 8-point rating scale, and their responses are then averaged into separate composite scores for commitment, satisfaction, investments, and alternatives. These composite scores are subsequently believed to be less saturated with measurement error and are entered into the regression model. The goals thus reflect primarily statistical concerns—estimating population parameters in the regression analysis from presumably continuous qualities—rather than the pursuit of causality via an integrated model. In the realm of observation oriented modeling, the issue of causality cannot be so easily avoided because the algorithms underlying most analyses depend on rotating the effect into conformity with the cause. Judging the first commitment and satisfaction items presented previously as final cause and effect, respectively, yields the following expression:

$$\text{Commitment} \rightarrow \text{satisfaction},$$

which is not equivalent to an expression that considers commitment as the effect:

$$\text{Satisfaction} \rightarrow \text{commitment}.$$

When analyzing these expressions in the Observation Oriented Modeling software, conforming satisfaction to commitment will not likely yield equivalent results compared to conforming commitment to satisfaction. Indeed, differences between causes and effects must be confronted at the level of the units of observations because of the asymmetric nature of the rotational procedures described in Chapter 2.

When discussing efficient cause, philosophers note that it does not follow from "*x* caused *y*" that "*y* caused *x*" (Dowe, 1992). Similarly, in the Observation Oriented Modeling software, the results obtained from rotating one set of ordered observations to another will not necessarily match those when switching and then rotating the ordered observations. The practical implication of this asymmetry is that when one attempts to bring a small number of units of observations into conformity with a larger number of units, ambiguous or incorrect classifications will be more likely to result. The use of observation oriented modeling therefore requires careful attention to the number of units posited for each set of observations.

This runs contrary to the habits developed within the Pearsonian–Fisherian tradition, in which differences in scale units are often ignored, or the scales are transformed (e.g., into *z* scores) or treated within the latent variable framework. With observation oriented modeling, one instead stays close to the observations as they are ordered in the integrated model and understood in terms of their causal structures and interrelationships.

The broader implications of this asymmetry are yet to be worked out, but at least in terms of efficient cause, there may be some precedent for expecting a conservation in the number of units between causally connected pairs of ordered observations. Discussions of causal asymmetry have been spurred by the revolution in physics instigated by Einstein and his contemporaries as well as the technological advances that have permitted physicists to test and explore their theories. For instance, Wesley Salmon (1998, see Chapter 17) discusses studies in quantum physics regarding the angular momentum between pairs of subatomic particles that were not possible before the construction of the Stern–Gerlach apparatus. Phil Dowe (1992) similarly relies on modern physics for supporting evidence for his conserved quantity theory of efficient cause.[19] In his theory, Dowe argues that both causal processes and causal interactions involve the conservation of quantities, such as mass, energy, and momentum. The history of objects can be represented as points in a two-dimensional space–time diagram. At particular intersections in the diagram, the histories of events may cross and impact each other in various ways, but in every case, quantities are conserved. For observation oriented modeling, one limitation of this work is its reliance on physics and the reduction of all processes to continuous quantities, but it nonetheless points to original thought based on a realist philosophy that might be built upon to entail discrete aspects of nature in discussions of asymmetry from the perspective of Aristotle's four causes.[20] Most important, careful consideration of causes and effects goes hand-in-hand with the development of integrated models that are necessary if psychologists are to create a *scientia naturalis* of persons.

[19] It is noteworthy that Salmon offers a general endorsement of Dowe's theory in the first chapter of his *Causality and Explanation* (Salmon, 1998).

[20] Dowe (1992b, pp. 214–215) discusses the reduction to physics. He also discusses other possible viewpoints that may fit better with the biological sciences and presumably psychology. Peter Menzies' paper, "Against Causal Reductionism" (1988), provides an excellent starting point for the philosophical discussion of causal reductionism as well as the assignment of causal powers to mental states.

This page intentionally left blank

CHAPTER 10

Coda

Contents

Introduction 219
Seven Principles of Observation Oriented Modeling 221
Summary 230

INTRODUCTION

> *Whatever I had admitted until now as most true I took in either from the senses or through the senses; however, I noticed that they sometimes deceived me. And it is a mark of prudence never to trust wholly in those things which have once deceived us.*
>
> —***René Descartes (1941/1979, p. 13)***

Thus wrote René Descartes in the 1630s as the initial move in his infamous method of doubt that eventually led to the detachment of the senses from the act of knowing:

> *But lo and behold, I have arrived on my own at the place I wanted. Since I know that bodies are not, properly speaking, perceived by the senses or by the faculty of imagination, but only by the intellect, and since, moreover, I know that they are not perceived by being touched or seen, but only insofar as they are expressly understood, nothing can be more easily and more evidently perceived by me than my mind.*
>
> ***(p. 23)***

The resemblance of this conclusion to Pearson's discussion of a "shadow table" (see Chapter 5) approximately 250 years later is certainly no coincidence because Descartes is the fount of Comte's positivism from which Pearson drew his inspiration. In positivism one finds a repudiation of philosophy and metaphysics and of all the traditional sciences save the methodology of the natural sciences. In short, one finds a dis-integration of the unity of the sciences sought by Thomas Aquinas. Gone is the notion that the sciences can be meaningfully integrated primarily on the basis of the act of knowing through the senses and secondarily on the methods and conceptual systems (acts of judgments) used by the persons studying nature. As mentioned in Chapter 1, positivism annihilates the subject–object dialectic, and Descartes laid the foundation for this destruction by going

Observation Oriented Modeling
ISBN 978-0-12-385194-9, Doi:10.1016/B978-0-12-385194-9.10010-6
© 2011 Elsevier Inc.
All rights reserved.

against a centuries-old anthropology that regarded each person as a composite of senses, passions, intellect, and will. For the Medieval scholar, a human being was regarded as a microcosm because it seemed all of the powers of the universe were wrapped up in each person. For Thomas Aquinas, "'Person' signifies what is most perfect in all nature—that is, a subsistent individual of a rational nature" (*Summa Theologica*, 1980, Ia, 29, 3). The human being further subsists as an ordered composite of matter and form, and Aquinas is stating that one of the powers that sets human persons apart from other creatures and things in nature is the capacity for intellection. Intellection is the process by which a person abstracts from the percept of a concrete, material object universal concepts through which the object can be understood (Wallace, 1996, pp. 131–135). This definition necessarily implicates the senses in the act of knowing, and its cogency rests on the validity of the subject–object dialectic. By concluding he can only be certain that he is "a thing that thinks," Descartes corrupts the integrated notion of man and sets the stage for the complete annihilation of the subject–object dialectic in later philosophical schools of thought.[1]

Against this backdrop, observation oriented modeling can be seen as a Gestalt shift from the positivism inherent in modern psychological and social science research to moderate realism.[2] In other words, observation oriented modeling represents a restoration of the subject–object dialectic and, with Aristotle and Aquinas, an acknowledgment that the process of understanding is rooted in the senses. This acknowledgment places the focus,

[1] Descartes only sets the stage because in his *Meditations* he attempts to recover the subject–object dialectic through an act of faith; specifically, the actuality of the material realm outside the mind cannot be demonstrated, but we can be certain of God's existence. Because God is a loving God, He would not deceive us with such powerful images of a material realm. Because God is no deceiver and we can trust Him, the material realm so powerfully present in one's mind does indeed exist. Immanuel Kant would later dispense with the centrality of the great metaphysical questions, and August Comte would dispense with God altogether and consequently the Cartesian justification for the entire subject–object dialectic. Cardinal Desiré Mercier, in *The Origins of Contemporary Psychology* (1918), commented on Descartes' reduction of man thusly: "In his study of the soul, he reduces all its activity to thought, and its nature to its aptitude for thinking. In his study of bodies, including that of the human body, he reduces all their properties to extension, all their activities to motion" (p. 27). Cardinal Mercier was also one of the first to trace this philosophy through Kant to psychology: "In fine, the Kantian *Critique* is a theory, deductive in form, of the positivism outlined by Comte and practiced by most contemporary psychologists" (p. 172).

[2] An example of an endorsement of realism in the social sciences comes from Manicas (2006).

then, on the things of study for any science and brings us to seven principles of observation oriented modeling presented in this final chapter. These principles are not given in any particular order, and they serve as summaries for the primary ideas and arguments found in preceding chapters.

SEVEN PRINCIPLES OF OBSERVATION ORIENTED MODELING

The first principle of observation oriented modeling is the primacy of the real. For the person who adopts observation oriented modeling, this means the beginning and ending point of all inquiry is the object of study. For psychologists, in particular, this means that the person, as a subsistent subject, is the beginning and end of every investigation. As argued in Chapter 6, psychologists wish to make general statements about persons, which is a proper goal of science, but through aggregate-based statistics framed in null hypothesis significance testing (NHST), only generalizations about population parameters are typically permitted. Observation oriented modeling encourages the psychologist to return the focus to the persons in the study, who are the objects of investigation.

In a more general sense, the shift from positivism and other forms of idealism to moderate realism can be understood by examining contemporary psychology textbooks. It is telling that most are organized around schools of thought, topics, or methods. For instance, introductory textbooks are organized into chapters on cognitive psychology, personality psychology, social psychology, clinical psychology, etc. Consistent with Descartes, such a scheme places emphasis on the conceptual systems through which persons are known rather than on the persons themselves. By comparison, an early textbook, *General Psychology*, written in 1937 by Robert Brennan, a Thomistic psychologist, was organized around the different acts, powers, and habits of persons. Brennan therefore organized his book into chapters on a person's senses, imagination, instincts, intellect, volition, and emotions. His book is akin to an operating manual for an automobile—a veritable manual for personhood. There can be no disagreement with the argument that modern psychology has fragmented into numerous subdomains. From the perspective of observation oriented modeling, only by recognizing the primacy of the subsistent person (an existing, integrated whole), who is the object of study, will psychology find hope in ever becoming unified.[3]

[3] Although he did not explicitly adopt a position of philosophical realism, Peter Molenaar (2004) recently published a manifesto on placing the person at the center of modeling.

Accompanying the primacy of the real is the second principle, the primacy of observations. For the modern psychologist and social scientist who relies on the Pearsonian–Fisherian framework, this principle requires that the notion of latent variable be abandoned altogether or at least employed in a much more restricted and judicious manner. The switch to observation oriented modeling also requires an honest examination of the numerous assumptions that enter into statistical modeling, particularly modern psychometric theory and NHST. As noted in Chapter 8, most forms of latent variable models suffer from score indeterminacy, which means that scores on the constructed latent variates cannot be unambiguously connected to the observations. The common response to this dilemma has been to ignore indeterminacy altogether or to treat the latent variates as somehow more real, or at least less error-laden, than the observations. The cost of this Cartesian bias—sacrificing the reality of observations for purely conceptual statistical models—is amplified when the numerous unwarranted assumptions underlying these models are examined. Classical models of psychological measurement, for instance, assume that errors are independent and normally distributed for each person being examined. As discussed in Chapter 8, this further requires the fantastic metaphorical assumption that each person's "brains are washed" between testing occasions. The estimation of *p* values in the NHST framework is, in most cases, also accompanied by numerous assumptions that are equally unlikely to hold in reality (see Chapter 5). Perhaps the unconscious hope throughout the years has been that the payoff for making these unwarranted assumptions would offset the costs, but the testimony of numerous psychologists and several organizations reviewed in Chapter 1 indicates otherwise.

Observation oriented modeling restores observations to their central role in scientific inquiry. The modeling techniques described in this book all work at the level of the observations for which the investigator must take greater responsibility. Consider the following Likert-type scale examined in Chapter 2:

Generally speaking, I like to spend most of my time alone.

Highly Disagree	*Disagree*	*Neither Disagree nor Agree*	*Agree*	*Highly Agree*
___	___	___	___	___

The observation oriented modeler does not presume that the quality (introversion) allegedly being "measured" by this item is structured as

a continuous quantity, and the observations obtained from this item are certainly not assumed to possess the requisite properties of continuous quantity (see Chapter 8). Without rigorous proof and demonstration, the assumption of continuous quantitative structure is not made, which is sufficient to disqualify most latent variable models in psychology today. Instead, the investigator who created the previous item takes full responsibility for how it orders the observations in the study. Why use only five units of observation rather than six or seven? Does the attribute or quality of interest admit to judgments of degree, such as "greater" or "lesser." Is the attribute bipolar in nature? Such questions are fundamental in observation oriented modeling because they will ultimately determine how the observations are to be ordered. Once observations are obtained for the item, they are analyzed "as is"—that is, as structured by the researcher—in a manner that is relatively free of assumptions. The binary Procrustes rotation algorithm and matching algorithms described in this book do not necessitate the invocation of latent variables or other linear composites, nor are they based on aggregates. The lone probability statistic, the chance value (or *c* value), is derived from a randomization test based on the obtained observations, and it requires only the assumption of complete independence to be valid. With such parsimony, observation oriented modeling permits the psychologist and social scientist to spend more time explaining the salient patterns in the units of observation (or lack thereof) than struggling with the indeterminacy of latent variables or suspect assumptions of statistical modeling in the Pearsonian–Fisherian tradition.

Free to focus on the ordered observations, the observation oriented modeler recognizes two additional principles: Aggregation often leads to obfuscation and outliers are people too. It has not been a tightly held secret that averaging numerical observations, particularly those obtained across people in a psychological study, can lead to obfuscation. As reviewed in Chapter 7 and demonstrated via example, working with aggregate data can even lead to the wrong conclusion being drawn. When aggregate statistics began to triumph as the way by which psychological and social science data would be both conceptualized and analyzed in the first half of the 20th century, comparative and behavioral psychologists exposed the many shortcomings of this approach. They advocated collecting numerous observations from only a few persons or a few organisms, analyzing the data using simple graphing techniques or mathematical models, and then generalizing their results to other persons or organisms via exact replications and descriptive equations. Single-subject designs were viewed as superior to

between-subject designs, and aggregation was considered as something to be generally avoid. These particular ideas are consistent with observation oriented modeling.[4]

On a deeper level, the arguments presented in Chapter 8 indicate that the computation of simple statistics such as means and variances may be suspect if the attributes or qualities under investigation are not structured as continuous quantities. More sophisticated techniques based on these simple aggregates, such as multiple regression, analysis of variance, and structural equation modeling, are consequently suspect as well. Even if legitimately computed, several important aggregate statistics (i.e., the mean and variance) are also sensitive to outliers, which are considered as nuisances in the Pearsonian–Fisherian tradition. Every student of psychology learns to struggle with the handling of outliers. Should outliers be discarded? Should the data be transformed? Should a Winsorized mean be computed? Should a nonparametric test be used instead of its parametric equivalent? By comparison, the techniques incorporated in observation oriented modeling are relatively insensitive to outliers. In the tradition of science discussed in Chapter 6, the student may consequently view outliers as opportunities to learn about the limitations of an integrated model or as vehicles of discovery rather than as nuisances.

Indeed, integrated models play a critical role in the principles that have previously been enumerated. The fifth principle of observation oriented modeling is therefore, think through the lens of an integrated model. From Aristotle's perspective, science is the process by which we develop certain knowledge of nature (*scientia naturalis*) through its causes. In more colloquial language, it is the process by which we come to confidently understand the structures and processes that constitute natural systems. An integrated model lays out, in a specific and schematic way, these causes—or structures and processes. As discussed in Chapter 9, by "certain knowledge," Aristotle meant deductively sound knowledge, particularly with regard to categorical syllogistic reasoning. With regard to "cause," Aristotle recognized four species of cause: formal, material, efficient, and final. It is by way of an integrated model that the

[4] An excellent text for single-subject designs is *Single Case Experimental Designs* by Barlow, Nock, and Hersen (2009). Although behaviorists and comparative psychologists are to be lauded for their attention to individual organisms or persons in their research, on the whole, they were unfortunately steeped in positivism. This philosophical position is of course contrary to the moderate realism of observation oriented modeling.

observation oriented modeler strives toward such certain knowledge of nature's causes.

William Wallace (1996) generically describes a model as follows:

> *For purposes here, a model will be taken to be an analogue or analogy that assists or promotes the gradual understanding of something not readily grasped in sense experience. Knowledge growth in humans is far from being a simple accumulation; rather it is a complex process whereby we perceive and learn from things around us by noting their similarities and differences.*
>
> ***(p. 5)***

An integrated model is similarly an analogy for existing things as they interact, move about, and change through time. It is also a more formal and disciplined mode of thought that constitutes the normal, everyday process of knowing. Imagine a psychologist working in her laboratory when two potential participants enter through the door. From the position of moderate realism, the psychologist immediately enters into an existential encounter with two other beings, which can be described as a subject (psychologist)–object (participants) dialectic. Through her senses, the researcher perceives the two beings in her presence and through her intellect recognizes them as female persons, which in this instance she finds surprising given that she only recruited males to participate in her study. "Person" and "female" are species concepts the psychologist abstracts via her intellect from the two subsistent beings in her presence. They are concepts that inhere in the two participants themselves and therefore exist both in her mind and outside her mind. They are what Wallace describes as "observable concepts" (pp. 254–257). The psychologist also recognizes two discrete, countable persons walking through the door when she expected only one participant. Here, she is using a metrical concept that is again both mental and extra-mental because it exists in her mind but is circumscribed by the subsistent beings (discrete, ordered wholes) in her presence. Finally, the psychologist may think to herself, "Well, I have two wonderful students willing to participate in my study today, but unfortunately this is an 'empty set' for my research because they are both females and I am testing only males," thus moving to a higher realm of abstraction with the concept of empty set. Such logical concepts are not tied to the existence of particular things; in other words, their foundation is not in the realm of the senses. Regardless of the level of abstraction, however, the moderate realist psychologist understands that "person," "female," and "two" are concepts abstracted by her intellect to help her understand the beings whom she

sensed walked through the door. If she were to somehow organize these terms on a piece of paper into a diagrammatic figure, perhaps supplemented with sketched stick figures of the students, she would be working with a model that stands as an analogue to her actual experience of the two participants. Once constructed, she could evaluate such a model using logical terms, test its correspondence with observations, and add concepts to the model in order to improve her understanding of the two participants or other participants as well. The model is therefore a bridge between the intra- and extra-mental realms that can be used both deductively and inductively in the attainment of knowledge.

Similarly, integrated models are to be understood as analogues of the natural order. When building an integrated model, the psychologist or social scientist makes explicit the orderings of observations associated with the things, people, qualities, or events under investigation and how they might change over time. As a scientific model, it is fundamentally tied to sense experience, whether it be through direct careful observation or through the special experience of a controlled experiment. As noted in Chapter 6, proceeding to increasingly greater levels of abstraction while simultaneously abandoning empirical methodology (but not the senses or observable world) leads to the realms of philosophy and metaphysics, which are necessary for a more complete and accurate understanding of nature. For the natural scientist, however, an integrated model must always be rooted and evaluated in the sensible realm, in the orderings of observations. The model must also be open to incorporation of all four causes because only in this way will it capture the structures and processes envisioned by the critics of modern psychology. J. S. Mill's canons of logic, although useful in a pragmatic context, are simply too restrictive. Also, whereas Galileo's triumph is sometimes regarded as a triumph over final cause, even Francis Bacon (1620/1952), who considered Aristotle's rules of demonstration as perverted, seemed to realize that when speaking of persons, it is still a useful concept:

> *Yet, the human understanding, incapable of resting, seeks for something more intelligible. Thus, however, whilst aiming at further progress, it falls back to what is actually less advanced, namely, final causes; for they are clearly more allied to man's own nature, than the system of the universe, and from this source they have wonderfully corrupted philosophy.*
>
> ***(p. 111)***

Moreover, Aristotle did not equate final cause with purpose but, rather, understood it to also entail the normal end states of natural systems that

make science possible (see Wallace, 1983). In observation oriented modeling, then, one need not shy away from including final causes in an integrated model. Joseph Rychlak's (1988) logical learning theory is a prime example of how both formal and final causes can be incorporated into a psychological model.

It should also be noted that by using an integrated model, the scientist avoids being crushed by the Scylla of operationalism and the Charybdis of latent variables. As noted previously, for the psychologist, a model is composed of concepts, but most of these concepts have their basis in the objective realm—that is, in the things of study. This is the philosophical position of moderate realism that stands in contrast to operationalism. The latter is purely subjective in nature and in league with Stevens' four scales of measurement—attributes, qualities, and measurement are all defined purely on the basis of the rules that the scientist follows. The early criticism of operationalism is that with respect to intelligence, for instance, there are as many attributes named "intelligence" as there are methods for "measuring" intelligence. The advent of latent variables can be seen as a partial response to the subjectivist trap of operationalism, but as argued in Chapter 8 and noted previously, latent variables can only be understood from a purely subjectivist position as well. Observation oriented modeling assumes that the attributes or qualities under investigation, such as intelligence, extraversion, and depression, inhere in the persons themselves. When the modeler goes about representing these qualities in an integrated model, he or she must make a commitment to the nature of the quality. If it is considered to be structured as a continuous quantity, then it must be represented as such in the model; if it is a quality that admits to judgments of "greater" and "lesser," then it must be represented as such in the model, with the appropriate units of observations made explicit. Defining units of observation in observation oriented modeling is therefore not an exercise in operationalism because it is a process that extends far beyond the variable-based terminology of the Pearsonian–Fisherian tradition. Again, an integrated model is inherently a causal model, with the potential of incorporating all four of Aristotle's causes. It therefore makes explicit the structures of attributes and qualities, how they interact, and how they may change through time, all of which extend far beyond the subjective activities of operationally defining a construct or positing a latent, indeterminate variable.

Once created, an integrated model is evaluated primarily on the basis of accuracy, yielding the sixth principle of observation oriented modeling, the

primacy of accuracy and repeatability. The cracks in the foundation of the Pearsonian–Fisherian tradition, particularly NHST, became visible in the 1990s, although they began to form in the mid-20th century. The surveys showing that a majority of psychologists cannot properly define the *p* value, which is the primary criterion for judging results obtained in the framework of NHST, were followed by an emphasis on effect sizes (see Chapter 7). This new emphasis, however, has by all appearances done nothing to thwart the misunderstandings of NHST, and it appears that one arbitrary (some would say mindless) ritual revolving round the *p* value has been augmented by ritualistic judgments of effect sizes as "small," "medium," or "large" (Gigerenzer, 2004). The fundamental, philosophical problem is that these methods simply do not place the emphasis where it should be—on accuracy. With the Observation Oriented Modeling software, the percent correct classification (PCC) index provides a commonsense, easy-to-interpret and communicable criterion for judging the accuracy of an integrated model. It is not based on aggregate statistics but instead on the observations as they are ordered in the model. Validity is generally defined in the social sciences as the accuracy of measurement, and validity coefficients in the Pearsonian–Fisherian tradition are often presented as correlation coefficients or squared correlation coefficients. Either way, they do not convey accuracy in a manner that is readily graspable by the intellect. For instance, does a correlation of .50, representing 25% shared variance between a predictor and criterion, indicate a high degree of validity—that is, accuracy? It is difficult to say, which is why arbitrary conventions are always necessary for interpreting such effect sizes. Suppose that this same value converted to a PCC index of 15%. Such a result immediately conveys failure in the integrated model because most of the observations (85%) could not be correctly classified. This is the sense of accuracy familiar to most scientists and consumers of scientific research, and it goes hand-in-hand with the logical certainty of scientific knowledge espoused by Aristotle.

Even the neophyte is aware that the certainty of *scientia naturalis* is buttressed by replication. One of the first questions asked, if not the first question asked, by an investigator following an experiment is, Can the result be replicated? If the integrated model is particularly novel or yields a bold prediction, then this question becomes even more important. The most frequent error made by psychologists and social scientists regarding the *p* value in NHST is considering it to be an index of how likely a result will replicate in an independent study. For instance, if the observed *p* value is .05,

the error is to interpret this value as indicating that if the study were replicated 100 times, only 5 times would it fail. This error is deeply rooted and has its origin in Fisher's original creation of NHST (see Chapter 5). The fact is that the *p* value is not in any way a replication coefficient, nor is the *c* value in observation oriented modeling. The latter only indicates the unusualness of the observed pattern of observations compared to randomized versions of the same observations. The generalizability of a result from a particular set of observations must therefore be won the old fashion way, via direct replication. Of course, an integrated model, as a causal model, carries with it the justification for generalization, but as stated previously, it must be rooted in the sensible world and yield repeatable, accurate judgments of observations.

A proper understanding of NHST and the *p* value leads to the final principle; estimate population parameters in their proper, limited role. As discussed in Chapters 5 and 6, the primary goal of NHST is to deliver a point estimate or interval estimate (i.e., a confidence interval) for a population parameter. These parameters are often aggregate statistics, such as the population mean, variance, or proportion. There can be no doubt that in some instances the estimation of population parameters is a goal worth pursuing. The clearest example is with election polling; what makes it clear is that the population is precisely defined and the sampling methods can be properly evaluated. In a U.S. presidential election, for instance, the population is composed of all registered voters—a clearly defined and empirically valid population. The exit polling sampling methods are not purely random, but they are usually stratified by counties within each state to ensure that the obtained sample is representative of the population. The population parameters are the proportions of voters casting their ballots for the different candidates and are themselves straightforward. NHST is clearly suited for the goal of predicting the winner of the election well before all of the ballots have been counted, and the history of the accuracy for which these predictions have been made supports this claim. In observation oriented modeling, however, the focus is on the causal, integrated model in which the estimation of population parameters will not play a central role. The goal is instead to create a model that will yield an impressively high PCC index upon repeated collections of observations. In other words, the goal is to create an integrated model that accurately and consistently explains the structures and processes that constitute a natural system. The goal is *scientia naturalis*, not the estimation of abstract aggregate parameters that often have no basis in sensible reality.

SUMMARY

The seven principles of observation oriented modeling are as follows:

- The primacy of the real.
- The primacy of observations.
- Aggregation often leads to obfuscation.
- Outliers are people too.
- Think through the lens of an integrated model.
- The primacy of accuracy and repeatability.
- Estimate population parameters in their proper, limited role.

Applying these principles will require patience and a willingness to question more than 70 years of methodological tradition in psychology and the other social sciences. It will require an innovative spirit as well to develop novel uses of set theory, graph theory, Boolean algebra, combinatorics, logical analysis, and other methods in the construction and evaluation of integrated models. On a more abstract level, the observation oriented modeler must be willing to question the Cartesian idealism inherent in reigning methodologies. This book represents a first step in the move from this idealism to a philosophical position that is much older and that has been tested, refined, and ameliorated by time. The overarching theme of Thomas Aquinas' writings is unity—the integration of metaphysics, philosophy, and natural science through the integration of that enigmatic microcosm, the human person. Perhaps by striking out in a new direction, one rooted in 2000 years of continuous thought, his vision can be realized.

REFERENCES

Adler, M. (1941). Introduction. In O. P. Robert Brennan. (Ed.), *Thomistic psychology* (pp. vii–xiv). New York: Macmillan.

Anderson, N. H. (1961). Scales and statistics: Parametric and nonparametric. *Psychological Bulletin, 58*, 305–316.

Aquinas, St. Thomas. (1963). *The division and methods of the sciences*. Toronto: Pontifical Institute of Mediaeval Studies.

Aquinas, St. Thomas. (1975). *Summa contra gentiles, Book III*. South Bend, IN: University of Notre Dame Press.

Aquinas, St. Thomas (2008). *Summa theologica*. Translation by the Fathers of the English Dominican Province. Retrieved from http://www.newadvent.org/summa/index.html.

Aristotle. (1971). *The complete works of Aristotle, Volumes 1 and 2*. Princeton, NJ: Princeton University Press.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 89–195). New York: Academic Press.

Bacon, F. (1952). Novum Organum. In M. Adler (Ed.), *Great books of the Western world (Vol. 30)*. Chicago: Encyclopedia Britannica. Original work published 1620.

Bakan, D. (1967). *On method: Toward a reconstruction of psychological investigation*. San Francisco: Jossey-Bass.

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs* (3rd ed.). Boston: Pearson.

Baron, R., & Kenny, D. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182.

Barrett, P. (2008). The consequence of sustaining a pathology: Scientific stagnation—A commentary on the target article "Is psychometrics a pathological science?" by Joel Michell. *Measurement, 6*, 78–83.

Barrett, P. (2009, July). *Interrater reliability (IRR): Measuring agreement, and nothing else*. Paper presented at the annual meeting of the International Society for the Study of Individual Differences, Evanston, IL.

Barrett, P., & Rolland, J. (2009). *The meta-analytic correlation between the Big Five personality constructs of emotional stability and conscientiousness: Something is not quite right in the woodshed*. Whitepaper retrieved from http://www.pbarrett.net (January 1, 2010).

Benedict XVI (2006, September). *Faith reason and the university: Memories and Reflections*. Retrieved from http://www.vatican.va/holy_father/benedict_xvi/speeches/2006/september/documents/hf_ben-xvi_spe_20060912_university-regensburg_en.html.

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research, 44*, 175–184.

Bhaskar, R. (1997). *A realist theory of science* (2nd ed.). London: Verso.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53*, 605–634.

Borsboom, D. (2008). Latent variable theory. *Measurement, 6*, 25–53.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*, 203–219.

Boynton, H. (1948). *The beginnings of modern science*. Roslyn, NY: Classics Club.

Brennan, R. (1937). *General psychology*. New York: Macmillan.

Brennan, R. (1941). *Thomistic psychology*. New York: Macmillan.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.
Burisch, M. (1984). Approaches to personality inventory construction. *American Psychologist, 39*, 214–227.
Burke, D. J. (1953). Additive scales and statistics. *Psychological Review, 60*, 73–75.
Butt, Z., Wagner, L., Beaumont, J., Paice, J., Peterman, A., Shevrin, D., et al. (2008). Use of a single-item screening tool to detect clinically significant fatigue, pain, distress, and anorexia in ambulatory cancer practice. *Journal of Pain and Symptom Management, 35*, 20–30.
Carnahan, C., & McFarland, S. (2007). Revisiting the Stanford Prison Experiment: Could participant self-selection have led to the cruelty? *Personality and Social Psychology Bulletin, 33*, 603–614.
Cervone, D. (2004). The architecture of personality. *Psychological Review, 111*, 183–204.
Chrisman, N. R. (1998). Rethinking levels of measurement for cartography. *Cartography and Geographic Information Systems, 25*, 231–242.
Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science, 3*, 186–190.
Cohen, B. (1985). *Revolution in science.* Cambridge, MA: Harvard University Press.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Routledge. Original work published 1969.
Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997–1003.
Collins, L., Graham, J., & Flaherty, B. (1998). An alternative framework for defining mediation. *Multivariate Behavioral Research, 33*, 295–312.
Connor, C. (2005). *A people's history of science.* New York: Nation Books.
Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Chicago: Rand McNally.
Crombie, A. C. (1961)*Augustine to Galileo: Science in the later Middle Ages and early modern times (Vol. 2)*. Cambridge, MA: Harvard University Press.
Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*, 116–127.
Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
Crowley, C. O. P. (1996). *Aristotelian–Thomistic philosophy of measure and the International System of Units (SI)*. Lanham, MD: University Press of America.
Cudeck, R., & MacCallum, R. (2007). *Factor analysis at 100: Historical developments and future directions*. Mahwah, NJ: Erlbaum.
Danziger, K. (1990). *Constructing the subject*. Cambridge, UK: Cambridge University Press.
Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology, 8*, 377–383.
Descartes, R. (1979). *Meditations on first philosophy.* Indianapolis, IN: Hackett. Original work published 1641.
Dougherty, J. (2009). *Wretched Aristotle.* New York: Lexington Books.
Dowe, P. (1992a). Process causality and asymmetry. *Erkenntnis, 37*, 179–196.
Dowe, P. (1992b). Wesley Salmon's process theory of causality and the conserved quantity theory. *Philosophy of Science, 59*, 195–216.
Dozois, D., Bieling, P., Patelis-Siotis, I., Hoar, L., Chudzik, S., McCabe, K., et al. (2009). Changes in self-schema structure in cognitive therapy for major depressive disorder: A randomized clinical trial. *Journal of Consulting and Clinical Psychology,* 77, 1078–1088.
Duhem, P. (1954). *The aim and structure of physical theory.* Princeton, NJ: Princeton University Press. Original work published 1906.
Duhem, P. (1985). *Medieval cosmology: Theories of infinity, place, time, void, and the plurality of worlds*. Chicago: University of Chicago Press.

Edgington, E., & Onghena, P. (2007). *Randomization tests* (4th ed.). London/Boca Raton, FL: Chapman & Hall/CRC Press.
Erceg-Hurn, D., & Mirosevich, V. (2008). Modern robust statistical methods. *American Psychologist, 63*, 591–601.
Feldman Barrett, L. (2005). How random is that? *APS Observer, 18*(9).
Fisher, R. A. (1929). Letter to Major Leonard Darwin. Retrieved September 12, 2008, from http://post.queensu.ca/~forsdyke/bateson3.htm
Fisher, R. A. (1930). *Statistical methods for research workers* (3rd ed.). Edinburgh, UK: Oliver & Boyd.
Fisher, R. A. (1937). *The design of experiments* (2nd ed.). Edinburgh, UK: Oliver & Boyd.
Fitzsimons, G. (2008). Death to dichotomizing. *Journal of Consumer Research, 35*, 5–8.
Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin, 87*, 564–567.
Gardner, M. (1970, October). Mathematical games: The fantastic combinations of John Conway's new solitaire game "Life." *Scientific American, 223*, 120–123.
Gelfand, L., Mensinger, J., & Tenhave, T. (2009). Mediation analysis: A retrospective snapshot of practice and more recent directions. *Journal of General Psychology, 136*, 153–176.
Gerrig, R. J., & Zimbardo, P. G. (2002). *Psychology and life* (16th ed.). Boston: Allyn & Bacon.
Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics, 33*, 587–606.
Gigerenzer, G. (2009). Surrogates for theory. *APS Observer, 22*, 21–23.
Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989). *The empire of chance*. Cambridge, UK: Cambridge University Press.
Gilbert, W. (1952). On the loadstone and magnetic bodies. In M. Adler (Ed.), *Great books of the Western world (Vol. 28)*. Chicago: Encyclopedia Britannica. Original work published 1600.
Gilson, E. (1990). *Methodical realism*. Front Royal, VA: Christendom Press.
Green, B. F. (1952). The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika, 17*, 429–440.
Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and *p* values: What should be reported and what should be replicated? *Psychophysiology, 33*, 175–183.
Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6*, 430–450.
Grice, J. W., Krohn, E. J., & Logerquist, S. (1999). Cross-validation of the WISC-III factor structure in two samples of children with learning disabilities. *Journal of Psychoeducational Assessment, 17*, 236–248.
Harper, T. (1940). *The metaphysics of the school (3 vols.)*. New York: Peter Smith.
Harré, R. (1970). *The principles of scientific thinking*. Chicago: University of Chicago Press.
Harré, R. (1986). *Varieties of realism: A rationale for the natural sciences*. Hoboken, NJ: Blackwell.
Harré, R., & Secord, P. (1973). *The explanation of social behavior*. Totowa, NJ: Littlefield, Adams.
Harris, R. J. (1994). *Analysis of variance primer*. Itasca, IL: Peacock.
Harris, R. J. (1997). Significance tests have their place. *Psychological Science, 8*, 8–11.
Hartl, D., & Fairbanks, D. (2007). Mud sticks: On the alleged falsification of Mendel's data. *Genetics, 175*, 975–979.
Hartl, D., & Orel, V. (1992). What did Gregor Mendel think he discovered? *Genetics, 131*, 245–253.
Harvey, W. (1952). Anatomical exercises on the generation of animals. In M. Adler (Ed.), *Great books of the Western world (Vol. 28)*. Chicago: Encyclopedia Britannica. Original work published 1651.

Hemphill, J. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58*, 78–79.
Hock, R. R. (2009). *Forty studies that changed psychology: Explorations into the history of psychological research* (6th ed.). Upper Saddle River, NJ: Prentice Hall.
Howell, D. C. (2007). *Statistical methods for psychology* (6th ed.). Belmont, CA: Thomson-Wadsworth.
Hurley, J. R., & Cattell, R. B. (1962). The Procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioral Science, 7*, 258–262.
Jaki, S. L. (1966). *The relevance of physics*. Chicago: University of Chicago Press.
Jaki, S. L. (1978). *The origin of science and the science of its origin*. South Bend, IN: Regnery/Gateway.
Jaki, S. L. (2006). *A late awakening and other essays*. Port Huron, MI: Real View Books.
Johnson, H. M. (1936). Pseudo-mathematics in the mental and social sciences. *American Journal of Psychology, 48*, 342–351.
Kassing, L. R. (1997). Interpersonal relationships: Distinguishing between cyclical and noncyclical romantic relationships (unpublished masters thesis). Edwardsville: Southern Illinois University.
Kazdin, A. (2007). Mediators and mechanisms of change in psychotherapy research. *Annual Review of Clinical Psychology, 3*, 1–27.
Kelly, G. A. (1955). *The psychology of personal constructs*. New York: Norton.
Kennedy, D. (1907). Albertus Magnus. In *The Catholic encyclopedia*. New York: Robert Appleton Company. Retrieved from www.newadvent.org/cathen/01264a.htm.
Keselman, H., Algina, J., Lix, L., Wilcox, R., & Deering, K. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods, 13*, 110–129.
Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science, 16*, 345–352.
Kuhn, T. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.
Lakatos, I., & Musgrave, A. (1979). *Criticism and the growth of knowledge*. Cambridge, UK: Cambridge University Press.
Lamiell, J. T. (1997). Individuals and the differences between them. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 118–141). San Diego: Academic Press.
Lamiell, J. T. (2003). *Beyond individual and group differences: Human individuality, scientific psychology, and William Stern's critical personalism*. Thousand Oaks, CA: Sage.
Lamiell, J. T. (2009). The characterization of persons: Some fundamental conceptual issues. In P. Corr, & G. Matthews (Eds.), *The Cambridge handbook of personality psychology* (pp. 72–86). Cambridge, UK: Cambridge University Press.
Latané, B., & Darley, J. (1970). *The unresponsive bystander: Why doesn't he help?* New York: Appleton-Century-Crofts.
Lefebvre, V. A. (2001). Algebra of conscience. In W. Leinfellner, & G. Eberlein (Eds.), *Philosophy and methodology of the social sciences (Vol. 30)*. Boston: Kluwer.
LeLorier, J., Genevieve, G., Benhaddad, A., Lapierre, J., & Derderian, F. (1997). Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *New England Journal of Medicine, 337*, 536–542.
Leplin, J. (1984). *Scientific Realism*. Berkeley, CA: University of California Press.
Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 1–55.
Locksley, A., Ortiz, V., & Hepburn, C. (1980). Social categorization and discriminatory behavior: Extinguishing the minimal intergroup discrimination effect. *Journal of Personality and Social Psychology, 39*, 773–783.

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8*, 750–751.

Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*, 1–27.

Lykken D. (1991). What's wrong with psychology anyway? In D. Cicchetti & W. Grove (Eds.), *Thinking clearly about psychology* (Vol. 1, pp. 61–87). Minneapolis: University of Minnesota Press.

MacKinnon, D. (2008). *Introduction to statistical mediation analysis*. New York: Erlbaum.

Manicas, P. (2006). *A realist philosophy of social science*. Cambridge, UK: Cambridge University Press.

Manning, R., Levine, M., & Collins, A. (2007). The Kitty Genovese murder and the social psychology of helping. *American Psychologist, 62*, 555–562.

Maraun, M. D. (1996a). Metaphor taken as math: Indeterminacy in the factor analysis model. *Multivariate Behavioral Research, 31*, 517–538.

Maraun, M. D. (1996b). The claims of factor analysis. *Multivariate Behavioral Research, 31*, 673–689.

Maraun, M. D., & Grabriel, S. (2010). Killeen's (2005) p_{rep} coefficient: Logical and mathematical problems. *Psychological Methods, 15*, 182–191.

Maraun, M. D., & Halpin, P. (2008). Manifest and latent variates. *Measurement, 6*, 113–117.

Maraun, M. D., Slaney, K., & Gabriel, S. (2009). The Augustinian methodological family of psychology. *New Ideas in Psychology, 27*, 148–162.

Maritain, J. (1951). *Philosophy of nature*. New York: Philosophical Library.

Marsh, H., Hau, K., & Zhonglin, W. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*, 320–341.

Meehl, P. E. (1973). *Psychodiagnosis: Selected papers*. Minneapolis: University of Minnesota Press.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806–834.

Mendel, G. (1964). Experiments in plant hybridization. In J. H. Bennett (Ed.), *Experiments in plant hybridization; Mendel's original paper in English translation, with commentary and assessment by Sir Ronald A. Fishes, together with a reprint of W. Bateson's biographical notice of Mendel*. Edinburgh, UK: Oliver & Boyd.

Menzies, P. (1988). Against causal reductionism. *Mind, 98*, 551–574.

Merbitz, C., Morris, J., & Grip, J. (1989). Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation, 70*, 308–312.

Mercier, D. (1918). *The origins of contemporary psychology*. New York: Kenedy & Sons.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*, 355–383.

Michell, J. (1999). *Measurement in psychology: Critical history of a methodological concept*. Cambridge, UK: Cambridge University Press.

Michell, J. (2008). Is psychometrics pathological science? *Measurement, 6*, 7–24.

Michell, J., & Ernst, C. (1996a). The axioms of quantity and the theory of measurement, Part I: An English translation of Hölder (1901). *Journal of Mathematical Psychology, 40*, 235–252.

Michell, J., & Ernst, C. (1996b). The axioms of quantity and the theory of measurement, Part II: An English translation of Hölder (1901). *Journal of Mathematical Psychology, 41*, 345–356.

Mischel, W. (2004). Toward an integrative science of the person. *Annual Review of Psychology, 55*, 1–22.
Molder, D., & Dweck, C. (2001). Finding meaning in psychology: A lay theories approach to self-regulation, social perception, and social development. *American Psychologist, 61*, 192–203.
Molenaar, P. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2*, 201–218.
Moleski, M. (2000). *Personal Catholicism: The theological epistemologies of John Henry Newman and Michael Polanyi*. Washington, DC: Catholic University Press.
Monaghan, F., & Corcos, A. (1985). Chi-square and Mendel's experiments: Where's the bias? *Journal of Heredity, 76*, 307–309.
Mulaik, S. (1972). *The foundations of factor analysis*. New York: McGraw-Hill.
Nagy, M. S. (2002). Using a single-item approach to measure facet job satisfaction. *Journal of Occupational and Organizational Psychology, 75*, 77–86.
Naumann, L., Vazire, S., Rentfrow, P., & Gosling, S. (2009). Personality judgments based on physical appearance. *Personality and Social Psychology Bulletin, 35*, 1661–1671.
Newman, J. H. (1947). *A grammar of assent*. New York: Longmans, Green. Original work published 1870.
NIMH Consortium of Editors on Development and Psychopathology. (2000). Editorial statement. *Applied Developmental Science, 4*, 66.
Nogar, R. J. (1951). Toward a physical theory. *The New Scholasticism, 25*, 397–438.
Norris H. & Makhlouf-Norris F. (1976). The measurement of self-identity. In P. Slater (Ed.), *The measurement of intrapersonal space by grid technique: Explorations of intrapersonal space* (Vol. 1, pp. 79–83). London: Wiley.
Nunnally, J. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
Pace, E. (1912). Universities. In *The Catholic encyclopedia*. New York: Robert Appleton Company. Retrieved December 24, 2009, from http://www.newadvent.org/cathen/15188a.htm.
Pascual-Leone, A. (2009). Dynamic emotional processing in experiential therapy: Two steps forward, one step back. *Journal of Consulting and Clinical Psychology, 77*, 113–126.
Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.
Pearson, K. (1957). *The grammar of science*. New York: Meridian. Original work published 1892.
Plomin, R., DeFries, J., McClearn, G., & McGuffin, P. (2008). *Behavioral genetics*. New York: Worth.
Polanyi, M. (1968). Logic and psychology. *American Psychologist, 23*, 27–43.
Popper, K. (1979). Normal science and its dangers. In I. Lakatos, & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 52–58). Cambridge, UK: Cambridge University Press.
Popper, K. (1996). *The myth of framework: In defense of science and rationality*. New York: Routledge.
Porter, T. (2003). *Karl Pearson: The scientific life in a statistical age*. Princeton, NJ: Princeton University Press.
Prentice, D., & Miller, D. (1992). When small effects are impressive. *Psychological Bulletin, 112*, 160–164.
Redpath, P. (2010). Sign as a kind of opposition. *Semiotica*. in press.
Rizzi, A. (2004). *The science before science*. Baton Rouge, LA: IAP Press.
Robinson, D. (1986). *An intellectual history of psychology*. Madison: University of Wisconsin Press.
Rosenthal, R., & Rosnow, R. (2008). *Essentials of behavioral research*. Boston: McGraw-Hill. pp. 226–227.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences, 3*, 377–386.
Rosnow, R., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology, 57*, 221–237.
Rusbult, C. E. (1980). Commitment and satisfaction in romantic associations: A test of the investment model. *Journal of Experimental Social Psychology, 16*, 172–186.
Rychlak, J. F. (1981). *Introduction to personality and psychotherapy: A theory construction approach* (2nd ed.). Boston: Houghton Mifflin.
Rychlak, J. F. (1985). *A philosophy of science for personality theory* (2nd ed.). Malabar, FL: Krieger.
Rychlak, J. F. (1988). *The psychology of rigorous humanism* (2nd ed.). New York: New York University Press.
Salmon, W. C. (1998). *Causality and explanation*. New York: Oxford University Press.
Scargle, J. (2000). Publication bias: The "file drawer" problem in scientific inference. *Journal of Scientific Exploration, 14*, 91–106.
Schönemann, P. H. (1966). The generalized solution of the orthogonal Procrustes problem. *Psychometrika, 31*, 1–16.
Schönemann, P. H. (1994). Measurement: The reasonable ineffectiveness of mathematics in the social sciences. In I. Borg & P. Mohler (Eds.), *Trends and perspectives in empirical social research* (pp. 149–157). Berlin: de Gruyter.
Schönemann, P. H. (1996a). The psychopathology of factor indeterminacy. *Multivariate Behavioral Research, 31*, 571–577.
Schönemann, P. H. (1996b). Syllogisms of factor indeterminacy. *Multivariate Behavioral Research, 31*, 651–654.
Schönemann, P. H., & Scargle, J. D. (2008). A generalized publication bias model. *Chinese Journal of Psychology, 50*, 21–29.
Schwartz, R. M. (1997). Consider the simple screw: Cognitive science, quality improvement, and psychotherapy. *Journal of Consulting and Clinical Psychology, 65*, 970–983.
Schwartz, R. M., Reynolds, C. F., Thase, M. E., Frank, E., & Fasiczka, A. L. (2002). Optimal and normal affect balance in psychotherapy of major depression: Evaluation of the balanced states of mind model. *Behavioral and Cognitive Psychotherapy, 30*, 439–450.
Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.
Sidman, M. (1960). *Tactics of scientific research*. New York: Basic Books.
Steiger, J. (1996). Coming full circle in the history of factor indeterminacy. *Multivariate Behavioral Research, 31*, 617–630.
Steiger, J. H., & Schönemann, P. H. (1978). A history of factor indeterminacy. In S. Shye (Ed.), *Theory construction and data analysis*. Chicago: University of Chicago Press.
Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677–680.
Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.
Stevens, S. S. (1959). Measurement, psychophysics and utility. In C. W. Churchman & P. Ratoosh (Eds.), *Measurement: Definitions and theories* (pp. 18–63). New York: Wiley.
Sullivan, D. (2009). *An introduction to philosophy: Perennial principles of the classical realist tradition*. Charlotte, NC: TAN Books. Original work published 1957.
Thompson, B. (2004). The "significance" crisis in psychology and education. *Journal of Socio-Economics, 33*, 607–613.
Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory and Psychology, 19*, 579–599.
Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.

van den Berg, M., Timmermans, D., Knol, D., van Eijk, J., de Smit, D., van Vugt, J., et al. (2008). Understanding pregnant women's decision making concerning prenatal screening. *Health Psychology, 27*, 430–437.

Vazire, S., Naumann, L., Rentfrow, P., & Gosling, S. (2008). Portrait of a narcissist: Manifestations of narcissism in physical appearance. *Journal of Research in Personality, 42*, 1439–1447.

Velleman, P., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician, 47*, 65–72.

von Eye, A. (2002). *Configural frequency analysis: Methods, models, and applications*. London: Psychology Press.

Wain, H. M., Bruford, E. A., Lovering, R. C., Lush, M. J., Wright, M. W., & Povey, S. (2002). Guidelines for human gene nomenclature. *Genomics, 79*(4), 464–470.

Wallace, W. A. (1959). *The scientific methodology of Theodoric of Freiberg.* Fribourg, Switzerland: St. Paul's Press.

Wallace, W. A. (1972). *Causality and scientific explanation, Vol. I.* Ann Arbor, MI: University of Michigan Press.

Wallace, W. A. (1981). *Prelude to Galileo.* Boston: Kluwer.

Wallace, W. A. (1983). *From a realist point of view* (2nd ed.). Lanham, MD: University Press of America.

Wallace, W. A. (1996). *The modeling of nature.* Washington, DC: Catholic University Press.

Watson, R., & Evans, R. (1991). *The great psychologists* (5th ed.). New York: HarperCollins.

Weiling, F. (1986). What about R.A. Fisher's statement of the "too good" data of J.G. Mendel's Pisum paper? *Journal of Heredity, 77*, 281–283.

Weisberg, S. (1985). *Applied linear regression* (2nd ed.). New York: Wiley.

Wild, J. (1948). *Introduction to realistic philosophy.* New York: Harper & Row.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *The American Psychologist, 54*, 594–604.

Winch, R., & Campbell, D. (1969). Proof? No. Evidence? Yes. The significance of tests of significance. *The American Sociologist, 4*, 140–143.

Wolfram, S. (2002). *A new kind of science.* Champaign, IL: Wolfram.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practices, 52*, 33–45.

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; Measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation, 70*, 857–860.

Zuse, K. (1967). Rechnender raum. In *Elektronische Datenverarbeitung* (Vol. 8, pp. 336–344). Retrieved from http://www.idsia.ch/~juergen/digitalphysics.html.

Zuse, K. (1969). *Rechnender raum*. Braunschweig, Germany: Friedrich Vieweg & Sohn. English translation (February 1970). *Calculating Space*, MIT Technical Translation AZT-70-164-GEMIT. Cambridge, MA: Massachusetts Institute of Technology (Proj. MAC). Retrieved from http://www.idsia.ch/~juergen/digitalphysics.html.

INDEX

Adams, John Quincy, 143
Addition, deep structure, 22–23
Additivity, demonstration, 149–150
Adler, Mortimer J., 1
Analysis of variance (ANOVA), 144
ANOVA, see Analysis of variance
Aquinas, Thomas, 4, 6, 10, 49, 67, 92, 108, 151–152, 184, 201, 205, 207–209, 219–220, 230
Archimedes, 91
Aristotle, 4–7, 10–12, 20, 90, 92, 108, 148, 151, 154–155, 184–185, 201, 206–207, 215, 217, 220, 224, 226–227

Bacon, Francis, 226
Bacon, Roger, 92
Bakan, David, 2–3, 10
Barrett, Paul, 147, 153
Bateson, William, 80
Binary Procrustes rotation, 36–43
Box plot, 55
Brennan, Robert, 221
Bystander effect model, 94–99, 109

Causa, 11
Causality, 184–187, 214–217
Causation, 184–187
Cervone, Dan, 209
CFA, see Configural frequency analysis
Checklist, 165
Classification strength index (CSI), 51–52, 59, 110–114, 120, 162–163, 193–194, 199
Coefficient of determination, see Squared correlation
Cognitive–Affective Processing System, 208
Cohen, Jacob, 2, 10
Cohen's *d*, 129–131, 133–134, 138
Comte, Auguste, 67, 219
Configural frequency analysis (CFA), 199–200
Conforming matrix, 32
Correlation coefficient, 136–138
Cronbach, Lee, 82
Cronbach's alpha, 170
Crowley, Charles, 151–152
CSI, see Classification strength index
c value, 59–60, 110, 114, 116, 120, 141, 172, 174, 194, 198, 202, 204–205, 210, 214, 229

Danziger, Kurt, 72
Darley, John, 94–96, 99, 101–108, 112, 114, 117–118
Darwin, Charles, 66, 76–79
Data, see Deep structure
Deep structure
  addition, 22–23
  biological sex, 16
  division, 24–25
  ethnicity, 17
  Likert-type item, 14–15, 17, 24
  logical operators, 25–28
  matrices, 20–21
  modeling, see Modeling
  multiplication, 24–25
  overview, 9–10
  rotation
    binary Procrustes rotation, 36–43, 58
    Procrustes rotation, 29–36
  science origins, 10–12
  subtraction, 23–24
Descartes, René, 219–221
Division, deep structure, 24–25
DNA, 13
Dowe, Phil, 217
Duhem, Pierre, 11, 68, 70, 87

Efficient cause, 206–214
Einstein, Albert, 71
Error
  measurement, 169–175
  metaphysical error, 1–3
  methodological error, 1–3
  Types, 9, 63

Fechner, Gustav, 145
Final cause, 206–214
Fisher, Ronald, 7, 19, 66, 72–77, 79–82, 85, 88, 104–105, 184, 190, 203
Formal cause, 206–215
Furguson Committee, 146–147, 154

Galileo, 91, 226
Galton, Francis, 7, 66, 72, 82–83
G coefficient, 170
Gigerenzer, Gerd, 187
Gilbert, William, 91
Gilson, Etienne, 108
Gödel's theorem, 68
Grosseteste, 91–92

Halpin, Peter, 178
Harré, Rom, 1, 5, 108
Harris, Richard, 57–58
Hemphill, James, 134
Herclitus, 90
Hippocrates, 90
Hölder, Otto, 146, 148, 150, 153–154, 156, 161
Hume, David, 7, 190
Huxley, Thomas, 66
Hybridized null hypothesis significance testing, 84–85

ICC, see Intraclass correlation coefficient
Integrated model, 224–228
Intraclass correlation coefficient (ICC), 172, 174
IRT, see Item response theory
Item response theory (IRT), 176

Jaki, Stanley, 11

Kazdin, Alan, 183, 186
Kekulé, August, 13
Kelly, George, 164, 209–210, 214
Knowledge-and-Appraisal Personality Architecture model, 209
Kuhn, Thomas, 87–88, 106

Latané, Bibb, 94–96, 99, 101–108, 112, 114, 117–118
Latent variables, 175–181, 227
Lefebvre, Vladimir, 90, 97
Leibnitz, Gottfried, 91
Likert-type item, 14–15, 17, 24, 222
Logical hypothesis testing, 201–206
Luce, Duncan, 156
Lykken, David, 2

MacKinnon, David, 192
Magnus, Albertus, 91–92
Mann–Whitney *U* test, 144
Maraun, Michael, 178
Maritain, Jacques, 92, 104, 108
Measurement
  additive models, 166–168
  as knowing quantity, 151–158
  continuous quantitative measurement, 145–150
  error, 169–175
  latent variables, 175–181
  observation oriented modeling, 158–166
  overview, 143–145
Mediation models
  overview, 190–195
  person-centered mediation analysis, 195–201
Meehl, Paul, 2, 78–79, 81
Memory, classic model, 90
Mendel, Gregor, 79–81, 84, 180
Metaphysical error, 1–3
Methodological error, 1–3
Michell, Joel, 144–148, 153, 156, 162
Mill, John Stuart, 7, 12, 46, 190, 201, 226
Mischel, Walter, 208–209
Modeling
  additive models, 166–168
  aggregate testing, 104–108, 224
  bystander effect model, 94–99, 109
  deep structure, 45–63
  effect sizes
    aggregate effects, 122–123
    correlation and shared variance, 136–142
    differences between means, 123–136
  historical perspective, 87–94
  integrated model, 224–228
  mediation models

overview, 190–195
person-centered mediation analysis, 195–201
nature, 184–186
observation oriented modeling
overview, 108–120
measurements, 158–166
seven principles, 221–230
Molenaar, Peter, 221
Monopoly, 73
Mulaik, Stanley, 32
Multigram, 55–57, 62, 110, 114, 116, 164, 168, 193–195, 197, 200
Multiplication, deep structure, 24–25
Multi-unit frequency histogram, see Multigram

Newman, John Henry, 67–68, 70–71
Newton, Isaac, 91
Neyman, Jerzy, 85
NHST, see Null hypothesis significance testing
Null hypothesis, 48, 74
Null hypothesis significance testing (NHST)
aggregate testing, 105–107
American Psychological Association task force, 121, 123, 131, 142
overview and critique, 74–85, 94
population parameter estimation, 99–104, 117–119
p value estimation, 222, 228–229

Observation oriented modeling, see Modeling
Observed odds ratio (OR), 100–101
Odds ratio, see Observed odds ratio
OR, see Observed odds ratio
Ordered observation, 9

Path analysis model, 89–90
PCC, see Percent correct classification
PCT, see Personal construct theory
Pearson, Egon, 85
Pearson, Karl, 7, 19, 66–67, 69–73, 82–83, 85, 88, 104–105, 148, 184, 190, 203, 219
Pearsonian-Fisherian monopoly, 73
Pearson's chi-square, 144, 190
Percent correct classification (PCC), 53, 57–60, 109–110, 113–114, 116, 120, 140–142, 162–164, 167, 172, 193–194, 198, 200, 202, 204–206, 228–229
Personal construct theory (PCT), 164
Philosophical realism, 3–7
Plato, 70
Popper, Karl, 87, 186
Population correlation coefficient, 136
Precision, 19–20
Probability, statistics in deep structure modeling, 57–59
Procrustes rotation, 29–36, 223
p value, 59, 82, 117, 222, 228–229

Quantitas virtuis, 142, 153, 155–156

Randomized control experiment, 187–190
Repertory grid, 210–214
Robinson, Daniel, 1
Rotation, deep structure
binary Procrustes rotation, 36–43, 58
overview, 29
Procrustes rotation, 29–36
Rusbult, Caryl, 215
Rychlak, Joseph, 1, 12, 14, 208–209, 227

Salmon, Wesley, 217
Scatter plots, 159–160
Schönemann, Peter, 157, 179
Schwartz, Robert, 165
Science, origins, 10–12
Sidman, Murray, 116–117
Singular-value decomposition plot, 211
Spearman, Charles, 170
Spearman's correlation, 144
Squared correlation, 137–139, 166
Statistics, see also Modeling; Null hypothesis significance testing
historical perspective, 66–74
inferential statistics, 108–120
probability statistics in deep structure modeling, 57–59
Stern, William, 191
Stern–Gerlach apparatus, 217
Stevens, Stanley, 9, 145–146, 155, 190

Student's *t*-test, 76, 78, 128, 132, 144
Subtraction, deep structure, 23–24

Teleology, 208
Telosponse, 208
Theological science, 10
Thurstone, Louis, 145
Trendler, Günter, 157–158
t-test, 76, 78, 128, 132, 144
Tukey, John, 156

Validity, 228
Violence Risk Assessment Guide, 165
von Eye, Alexander, 199

Wallace, William, 108, 184–187, 201, 206–207, 220, 225, 227
Weschler Intelligence Scale for Children (WISC-III), 176–177
WISC-III, see Weschler Intelligence Scale for Children
Wright, Benjamin, 3

XOR, 27

Zimbardo, Philip, 189